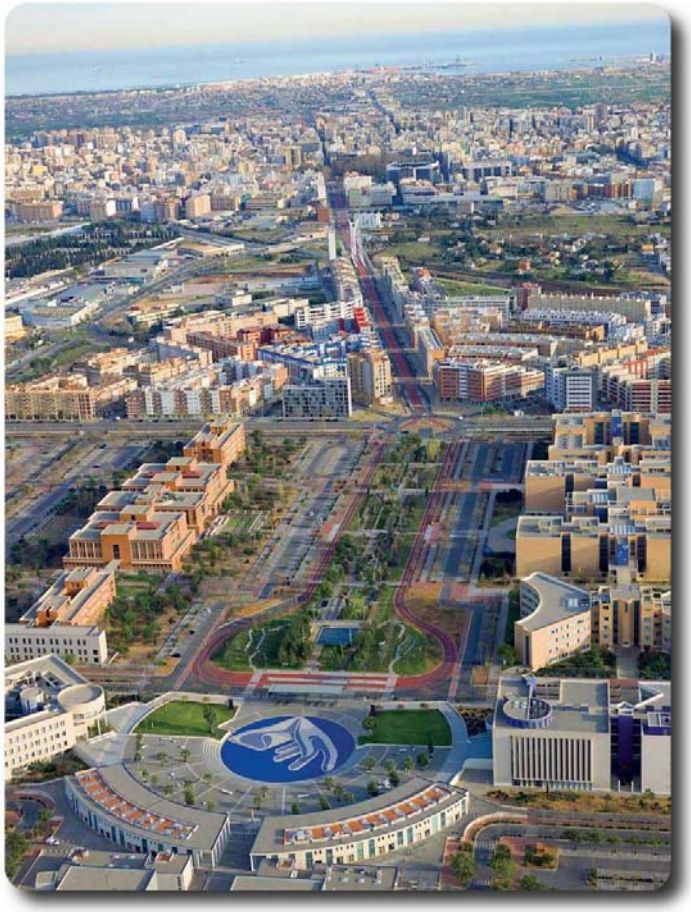
Connecting a Digital Europe through Location and Place

Selected best short papers and posters of the AGILE 2014 Conference
3-6 June 2014, Castellón, Spain



http://www.agile-online.org/index.php/conference/conference-2014

Editors: Joaquín Huerta, Sven Schade, Carlos Granell
ISBN: 978-90-816960-4-3
AGILE Digital Editions, June 2014

Sponsors


esri


ubik
Geospatial Solutions


FHC25
R + D SINCE 1996


esri España


Google


Geodan


prodevelop
Integración de tecnologías


52north
exploring horizons


Atos

Preface

Since 1998, the Association of Geographic Information Laboratories for Europe (AGILE) promotes academic teaching and research on geographic information at the European level. Its annual conference reflects the variety of topics, disciplines and actors in this research area. It provides a multidisciplinary forum for scientific knowledge production and dissemination and has gradually become the leading Geographic Information Science conference in Europe.

In addition to the conference full papers, which are published as a book by the Springer-Verlag, AGILE also encourages short paper and poster submissions for on-line publication. This year, 88 documents (70 short papers and 18 posters) were submitted under this option, of which 8 were re-submissions after the full paper proposals have been rejected. After a thorough selection and review process, a total of 44 short papers were approved for oral presentations, 17 short paper proposals were converted into posters, and 16 of the original 18 poster proposals were accepted. As much as we congratulate the authors for the quality of their work, we thank them for their contribution to the success of the AGILE conference and publication series. We also send our acknowledgements to the numerous reviewers for providing us with their thorough judgements. Their work was fundamental to select the very best contributions.

Under the sub-title *Connecting a Digital Europe through Location and Place*, these short paper and poster proceedings pay special attention to the role Geographic Information Science and Technology can play to connect European universities, research centres, industry, government and citizen in the digital information age. The scientific papers published in this volume cover a wide range of associated topics (from data capture to mapping and visualisation, and from human cognition to the political dimension), all organised following the sessions of the AGILE 2014 conference.
Organizing the program of an international conference and editing a proceedings of scientific papers takes time, effort and support. The input

from the AGILE Council and Committees was a tremendous asset for us and we are grateful to all members for their contributions.

We would also like to thank our sponsors (ESRI, Ubik Geospatial Solutions, FHC25, ESRI España, Google, Geodan, Prodevelop, 52North and AtoS), for their kind contributions to the 17th AGILE Conference on Geographic Information Science.

Castellón/Ispra, May 2014

Joaquín Huerta
Sven Schade
Carlos Granell

Programme Committee

Joaquín Huerta, University Jaume I of Castellón (Spain)
Sven Schade, European Commission- Joint Research Center (Italy)
Carlos Granell, European Commission- Joint Research Center (Italy)

Local Organising Committee

Laura Díaz, UNOPS (Spain)
Michael Gould, University Jaume I of Castellón (Spain)
Óscar Belmonte, University Jaume I of Castellón (Spain)
Marco Painho, Universidade Nova de Lisboa, ISEGI (Portugal)
Dori Apanewicz, University Jaume I of Castellón (Spain)
Ana Sanchis, University Jaume I of Castellón (Spain)
Luis Enrique Rodríguez, University Jaume I of Castellón (Spain)
Rubén Vidal, University Jaume I of Castellón (Spain)
Joaquín Torres-Sospedra, University Jaume I of Castellón (Spain)
Sven Casteleny, University Jaume I of Castellón (Spain)
Sergi Trilles, University Jaume I of Castellón (Spain)
Mauri Benedito, University Jaume I of Castellón (Spain)
Joan Pere Avariento, University Jaume I of Castellón (Spain)
Diego Gargallo, University Jaume I of Castellón (Spain)
David Rambla, University Jaume I of Castellón (Spain)

Scientific Committee

Peter Atkinson, University of Southampton (UK)
Fernando Bação, New University of Lisbon (Portugal)
Itzhak Benenson, Tel Aviv University (Israel)
Lars Bernard, TU Dresden (Germany)
Michela Bertolotto, University College Dublin (Ireland)
Ralf Bill, Rostock University (Germany)
Thomas Blaschke, University of Salzburg (Austria)
Thomas Brinkhoff, Jade University Oldenburg (Germany)
Bénédicte Bucher, IGN (France)
Gilberto Camara, National Institute for Space Research (Brazil)
Sven Casteleyn, University Jaume I of Castellón (Spain)
Nicholas Chrisman, University of Laval (Canada)
Christophe Claramunt, Naval Academy Research Institute (France)
Serena Coetzee, University of Pretoria (South Africa)
David Coleman, University of New Brunswick (Canada)
Lex Coomber, University of Leicester (UK)
Oscar Corcho, Universidad Politécnica de Madrid (Spain)
Max Craglia, European Commission-Joint Research Centre (Italy)
Joep Compvoets, KU Leuven Public Governance Institute (Belgium)
Anders Friis-Christensen, European Commission-Joint Research Centre (Italy)
Jérôme Gensel, University of Grenoble (France)
Yola Georgiadou, University of Twente (The Netherlands)
Michael Gould, University Jaume I of Castellón (Spain)
Carlos Granell, European Commission-Joint Research Centre (Italy)
Henning Sten Hansen, Aalborg University (Denmark)
Lars Harrie, Lund University (Sweden)
Francis Harvey, University of Minnesota (USA)
Gerard Heuvelink, Wageningen University (The Netherlands)
Stephen Hirtle, University of Pittsburgh (USA)
Hartwig Hochmair, University of Florida (USA)
Joaquín Huerta, University Jaume I of Castellón (Spain)

Bashkim Idrizi, State University of Tetova (Republic of Macedonia)
Mike Jackson, University of Nottingham (UK)Bin Jiang, University of Gävle (Sweden)
Didier Josselin, University of Avignon (France)
Derek Karssenberg, Utrecht University (The Netherlands)
Tomi Kauppinen, Aalto University (Finland)
Marinos Kavouras, National Technical University of Athens (Greece)
Karen Kemp, University of Southern California (USA)
Alexander Kotsev, European Commission-Joint Research Centre (Italy)
Menno-Jan Kraak, University of Twente (The Netherlands)
Patrick Laube, University of Zurich (Switzerland)
Robert Laurini, INSA-Lyon (France)
Francisco J Lopez-Pellicer, University of Zaragoza (Spain)
Joan Masó, CREAF (Spain)
Filipe Meneses, University of Minho (Portugal)
Peter Mooney, National University of Ireland Maynooth (Ireland)
Adriano Moreira, University of Minho (Portugal)
Jeremy Morley, University of Nottingham (UK)
Beniamino Murgante, University of Basilicata (Italy)
Pedro Muro Medrano, University of Zaragoza (Spain)
Javier Nogueras Iso, University of Zaragoza (Spain)
Toshihiro Osaragi, Tokyo Institute of Technology (Japan)
Frank Ostermann, University of Twente (The Netherlands)
Volker Paelke, Institute of Geomatics–Castelldefels (Spain)
Marco Painho, New University of Lisbon (Portugal)
Francesco Pantisano, European Commission-Joint Research Centre (Italy)
Poulicos Prastacos, Institute of Applied and Computational Mathematics FORTH (Greece)
Ross Purves, University of Zurich (Switzerland)
Francisco Ramos, University Jaume I of Castellón (Spain)
Martin Raubal, ETH Zurich (Switzerland)
Wolfgang Reinhardt, Universität der Bunderwehr Munich (Germany)
Femke Reitsma, University of Canterbury (New Zealand)
Claus Rinner, Ryerson University (Canada)
Jorge Rocha, University of Minho (Portugal)

Anne Ruas, IGN (France)
Maribel Yasmina Santos, University of Minho (Portugal)
Tapani Sarjakoski, Finnish Geodetic Institute (Finland)
Sven Schade, European Commission-Joint Research Centre (Italy)
Christoph Schlieder, University of Bamberg (Germany)
Monika Sester, Leibniz University Hannover (Germany)
Takeshi Shirabe, Royal Institute of Technology (Sweden)
Jantien Stoter, Delft University of Technology (The Netherlands)
Maguelonne Teisseire, IRSTEA (France)
Fred Toppen, Utrecht University (The Netherlands)
Joaquín Torres-Sospedra, University Jaume I of Castellón (Spain)
Nico Van de Weghe, Ghent University (Belgium)
Jos Van Orshoven, KU Leuven (Belgium)
Danny Vandenbroucke, KU Leuven (Belgium)
Lluis Vicens, Universitat de Girona (Spain)
Luis M. Vilches Blazquez, Universidad Politécnica de Madrid (Spain)
Agnès Voisard, FU Berlin and Fraunhofer FOKUS (Germany)
Monica Wachowicz, University of New Brunswick (Canada)
Robert Weibel, University of Zurich (Switzerland)
Stephan Winter, The University of Melbourne (Australia)
Mike Worboys, University of Maine (USA)
Bisheng Yang, Wuhan University (China)
Javier Zarazaga Soria, University of Zaragoza (Spain)
Alexander Zipf, Heidelberg University (Germany)

TABLE OF CONTENTS

Session: Analysis and Education

Geographic Information Technologies for analysing the digital footprint of tourists.
Toni Hernández, Josep Sitjar, Rosa Olivella and Lluis Vicens

"Troy is ours- How on earth could Clytaemnestra know so fast?".
Emmanuel Stefanakis and Titus Tienaah

Workforce Demand Assessment to Shape Future GI-Education- First Results of a Survey.
Barbara Hofer, Gudrun Wallentin, Christoph Traun and Josef Strobl

Some strategic national initiatives for the Swedish education in the geodata field.
Lars Harrie, Karin Larsson, David Tenenbaum, Milan Horemuz, Hanna Ridefelt, Gunnar Lysell, S. Anders Brandt, Eva A.U. Sahlin, Göran Adelsköld, Mats Högström and Jakob Lagerstedt

Session: Linked Data Web

Cadastral data integration through Linked Data.
Jhonny Saavedra Velasquez, Luis M. Vilches-Blázquez and Alberto Boada

GEOSUD SDI : Accessing Earth Observation data collections with semantic-based services.
Mathieu Kazmierski, Jean-Chistophe Desconnets, Bertrand Guerrero and Dominique Briand

Little Steps Towards Big Goals. Using Linked Data to Develop Next Generation Spatial Data Infrastructures (aka SDI 3.0).
Francis Harvey, Jim Jones, Simon Scheider, Adam Iwaniak, Iwona Kaczmarek, Jaromar Łukowicz and Marek Strzelecki

Session: Urban Dimension

Orchestrating the spatial planning process: from Business Process Management to 2nd generation Planning Support Systems.
Michele Campagna, Konstantin Ivanov and Pierangelo Massa

Recitoire: a tool for qualitative surveys involving citizens in urban planning projects.
David Noël, Marlène Villanova-Oliver and Jérôme Gensel

Planned vs. Real City: 3D GIS for Analyzing the Transformation of Urban Morphology.
Pilar Garcia-Almirall, Francesc Valls Dalmau and Montserrat Moix Bergada

Session: Qualitative Information

Qualitative Representation of Dynamic Attributes of Trajectories.
Tales Paiva Nogueira and Hervé Martin

VGI Edit History Reveals Data Trustworthiness and User Reputation.
Fausto D'Antonio, Paolo Fogliaroni and Tomi Kauppinen

A flexible framework for assessing the quality of crowdsourced data.
Sam Meek, Mike Jackson and Didier Leibovici

Session: Policy Dimension

Assessment of the integration of geographic information in e-government policy in Europe.
Glenn Vancauwenberghe, Danny Vandenbroucke, Joep Crompvoets, Francesco Pignatelli and Raymond Boguslawski

Publishing metadata of geospatial indicators as Linked Open Data: a policy-oriented approach.
Diederik Tirry, Ann Crabbé and Thérèse Steenberghen

Exploring the market potential for geo-ICT companies in relation to INSPIRE.
Glenn Vancauwenberghe, Piergiorgio Cipriano, Max Craglia, Cameron Easton, Giacomo Martirano and Danny Vandenbroucke

Session: Data Mining

Analysing spatiotemporal patterns of antibiotics prescriptions.
Luise Hutka and Lars Bernard

Influence of point cloud density on the results of automated Object-Based building extraction from ALS data.
Ivan Tomljenovic and Adam Rousell

Session: Routing

Street Network created by Proximity Graphs: Its Topological Structure and Travel Efficiency.
Toshihiro Osaragi and Yuko Hiraga

The effects of different verbal route instructions on spatial orientation.
Rui Li, Stefan Fuest and Angela Schwering

Session: Mapping and the Citizen Sensor COST TD1202

Semantic analysis of Citizen Sensing, Crowdsourcing and VGI.
Lex Comber, Sven Schade, Linda See, Peter Mooney and Giles Foody

Characteristics of Citizen-contributed Geographic Information.
Spyridon Spyratos, Michael Lutz and Francesco Pantisano

Is this Twitter Event a Disaster?.
André Dittrich and Christian Lucas

Session: Theory and Practice

A Geometric Configuration Ontology to Support Spatial Querying.
Kristin Stock

Spatiotemporal Data Complexity in electronic Airport Layout Plan and its visualization.
Shyam parhi

Towards Spatio-temporal Data Modeling of Geo-tagged Shipping Information.
Amin Mobasheri and Mohamed Bakillah

Towards Initiating Openlandmap Founded On Citizens' Science: The Current Status of Land Use Features of Openstreetmap In Europe.
Jamal Jokar Arsanjani, Eric Vaz, Mohamed Bakillah and Peter Mooney

Session: Observation Integration

Augmented Reality and GIS: On the Possibilities and Limits of Sensor-based AR.
Falko Schmid and Daniel Langerenken

Multi-sensory Integration for a Digital Earth Nervous System.
Frank Ostermann and Sven Schade

Session: Land Cover and Use COST TD1202

Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database.
Jacinto Estima, Cidália Fonte and Marco Painho

Cropland Capture: A Gaming Approach to Improve Global Land Cover.
Linda See, Tobias Sturn, Christoph Perger, Steffen Fritz, Ian McCallum and Carl Salk

Applying a CA-based model to explore land-use policy scenarios to contain sprawl in the case of Thessaloniki.
Apostolos Lagarias and Poulicos Prastacos

Session: Trajectories

Queues in Ski Resort Graphs: the Ski-Optim Model.
Tino Barras, Marut Doctor, Marc Revilloud, Michael Schumacher and Jean-Christophe Loubier

Real-time detection of anomalous paths through networks.
Steven Prager and R. Paul Wiegand

Session: Best Papers

Linking crowdsourced observations with INSPIRE.
Stefan Wiemann and Lars Bernard

Capability of movement features extracted from GPS trajectories for the classification of fine-grained behaviors.
Ali Soleymani, E. Emiel van Loon and Robert Weibel

Posters

Visualization of uncertain catchment boundaries and its influence on decision making.
Ulla Pyysalo and Juha Oksanen

SOS Server deployment for sharing environmental sensor data through the OTALEX-C Spatial Data Infrastructure.
Vivas White Pedro, Del Rey Amelia, Sanz Jorge and Brodin Ignacio

Design of the Data Transformation Architecture for the INSPIRE Data Model Browser.
Alberto Belussi, Piergiorgio Cipriano, Sara Migliorini, Mauro Negri and Giuseppe Pelagatti

Error propagation in a fuzzy logic spatial multi-criteria evaluation.
Lisa Bingham and Derek Karssenberg

ELF GeoLocator Service.
Pekka Latvala, Lassi Lehto and Jaakko Kähkönen

Enhancing the role of Citizen Sensors in Mapping: COST Action TD1202.
Giles Foody, Steffen Fritz, Linda See, Norman Kerle, Glen Hart and Cidalia Fonte

IDE-OTALEX C. The big challenge of the first Crossborder SDI between Spain and Portugal.
Teresa Batista, Carmen Caballero, Fernando Ceballos, Cristina Carriço, Pedro Vivas, José Cabezas, Luis Ferdández and Carlos Pinto-Gomes

Noise map: professional versus crowdsourced data.
Andrea Podör and András Révész

SDI strategic planning using the system dynamics technique: A case study in Tanzania.
Ali Mansourian, Alex Lubida, Ehsan Abdolmajidi, Petter Pilesjö and Micael Runnström

A conceptual representation for modelling the synchronization process of complex road networks.
Maria Dolores Arteaga Revert and Monica Wachowicz

Evaluation of subjective preferences regarding indoor maps: comparison of schematic maps and floor plans.
Luciene Delazari, Jeremy Morley and Suchith Anand

Using Crop Phenology to Assess Changes in Cultivated Land after the Anfal Genocide in Iraqi Kurdistan.
Lina Eklund, Andreas Persson, Jing Tang, Mitch Selander and Martin Borg

Fuzzy viewshed, probable viewshed, and their use in the analysis of prehistoric monuments placement in Western Slovakia.
Alexandra Rasova

Usability Patterns for Geoportals.
Christin Henzen and Lars Bernard

Agile access to sensor network.
Sergi Trilles, Óscar Belmonte Fernández, Laura Díaz Sánchez and Joaquín Huerta Guijarro

Utilization of NoSQL database for disaster preparedness.
Winhard Tampubolon

Latest Developments and activities in the Spanish NSDI.
Antonio Rodríguez

A Spatial Approach to Surveying Crime-problematic Areas at the Street Level.
Lucy Mburu and Alexander Zipf

It's GIrls' Day! What sketch maps show about girls' spatial knowledge.
Vanessa Joy Anacta and Thomas Bartoschek

3D Building Change Detection on the basis of Airborne Laser Scanning Data.
Karolina Korzeniowska and Norbert Pfeifer

Exploring twitter georeferenced data related to flood events: an initial approach.
Maria Antonia Brovelli, Carolina Arias Munoz, Giorgio Zamboni and Alexander Bonetti

Texas PanHandle Climate Change Interactive GIS Web Application.
Naga Raghuveer Modala

Data Scarcity or low Representativeness?: What hinders accuracy and precision of spatial interpolation of climate data?.
Avit Kumar Bhowmik and Ana Cristina Costa

Improving equity of public transportation planning. The case of Palma de Mallorca (Spain).
Maurici Ruiz, Joana Maria Seguí Pons, Jaume Mateu Lladó and Maria Rosa Martínez Reynés

The CartoCiudad gamble on Open Source and value-added services.
Alicia Gonzalez, Antonio Rodríguez, Celia Sevilla and Miguel Villalón

FOODIE: Farm-Oriented Open Data in Europe.
Miguel Ángel Esbrí

How Earth Observation, Crop Modeling, and ICT can help rice cultivation: the ERMES project.
Sven Casteleyn, Carlos Granell, Sergi Trilles, Joaquín Huerta, Mirco Boschetti, Lorenzo Busetto, Monica Pepe, Dimitrios Katsantonis, Roberto Confalonieri, Francesco Holecz, Javier García Haro and Ioannis Gitas

SHORT PAPERS

Session:
Data Capture and Mapping

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

A comparative study on VGI and professional noise data

Irene Garcia-Martí
GEOTEC
Research Group
University Jaume I, Spain
irene.garcia@uji.es

Joaquín Torres-Sospedra
GEOTEC
Research Group
University Jaume I, Spain
jtorres@uji.es

Luis E. Rodríguez-Pupo
GEOTEC
Research Group
University Jaume I, Spain
pupo@uji.es

Joaquín Huerta
GEOTEC
Research Group
University Jaume I, Spain
huerta@uji.es

Abstract

The ubiquitous nature of mobile devices and its growing presence in urban areas, turn them up into low cost environmental monitoring platforms. In this field, several authors made different efforts to provide alternatives to Sensor Networks, to assess noise pollution in cities using crowdsourcing techniques. In this sense, citizens might potentially produce large spatio-temporal datasets using their mobile devices to measure noise levels. There are few attempts of assessing the quality of the mobile noise samples on a real scenario and compare them to commercial data to evaluate if they are reliable enough. This contribution reviews the existing applications to collect or assess the quality of noise samples when they are used as sound level meters. Moreover, it presents the results of our experiment: the volunteer noise dataset generated in a 'mapping party' on our campus is compared to professional data. Results show that VGI data might be sufficient for multiple daily situations.

Keywords: Citizen Science, VGI, noise pollution, environmental monitoring, crowdsourcing, Smart Cities.

1 Introduction

According to United Nations report, the demographic growth over the next decades will be concentrated on cities and, by 2050, it is estimated that 70% of world people will be living in urban areas [18]. To improve economic and social conditions in urban environments, technicians and urban planners develop infrastructures that connect everyday living to the natural and informational resources to help making cities more sustainable [9, 17].

As population increases, cities become bigger and noisier and excessive noise levels have a direct impact on nature and environment: some species might be altering their behaviour to adapt to the increasing noise around us [4, 16, 1].

An effective way of monitoring our environment is through crowdsourcing [8]. With the application of this technique, we enable citizens to produce geographic information at a very low cost. Using VGI, we are capable of potentially extracting enough data from a city, with citizens' collaboration. However, due to the novelty of VGI approaches and its volunteer and collaborative nature, it is difficult to state if the quality of volunteer data is good enough for data analysis. To our knowledge, there are no approaches that compare commercial and volunteer noise data acquired through smartphones in a real scenario. Therefore, the work presented in this paper will try to provide a general overview of the potential of volunteer mobile noise monitoring.

Section 2 contains the related work, where we review several approaches considering noise monitoring and noise quality. In Section 3 we describe our data collections and the process carried out for the spatial analysis to obtain noise maps from our volunteer data. Section 4 presents the limitations and issues found in this project, whereas Section 5 outlines possible future lines of work. Finally, Section 6 summarizes the findings of our work.

2 Related work

Monitoring our environment is a crucial task to know how human activities affect our planet. In [5] Goodchild proposes a new way of acquiring environmental data and presents the concept of "citizens-as-sensors", also known in literature as crowdsourcing. There are several approaches to monitor noise pollution in urban environments applying crowdsourcing techniques. NoiseSpy [10] is a web platform that allows the measurement and real-time visualization of noise samples that the community of users uploads to a central server. NoiseTube [11] allows the creation of noise maps by sharing public measurements. This application provides to each user their personal exposure to noise pollution. NoiseBattle [6, 7] is a gamified application for noise sampling that tackles the problem of motivation and engagement of users for environmental monitoring.

However, these applications are focused on the noise collection, information visualization or user engagement and motivation, but do not consider directly the goodness of noise data mobile devices can acquire on a real scenario. In [13] it is possible to find the demonstration on how a mobile device after a calibration process can produce highly accurate measurements, when compared to a professional device. Similarly, [12] suggest that it is possible to obtain with mobile devices data with a precision and quality just few decibels different from professional devices.

In [3, 15] it is possible to find a discussion about how good might be the quality of noise samples collected with mobile phones. The article describes an experiment where three mobile devices and a sound level meter are exposed in few tests to different sources of noise.

3 Producing noise maps

This section explains the process followed to obtain noise maps to compare VGI and professional datasets. First, we describe both data collections, then we explain the process carried out during the analysis part and finally we present our results separately: on one hand all volunteer data involved in the project and, on the other, in particular for two concrete types of mobile devices.

3.1 Data collection description

As stated before, two different noise data sets were used: one collected by volunteers and the other one obtained with professional means.

Crowdsourced data. Volunteer dataset was obtained helding a 'mapping party' on University Jaume I Campus with members of GEOTEC Group. Data were collected in the central part of campus, comprising three faculties, the access gate and the central garden. In total, the study area is 585 meters long and 487 meters wide, giving an area of 0'285 km^2. Within this area, participants were encouraged to take measurements in 30 predefined locations.

The software used to collect noise samples is described in [6, 7]. An extra layer with the 30 points of the grid was added to help the user taking the measurement in the recommended locations. The experiment was repeated on the same places as the professional company did in order to generate maps that were possible to compare. Weather conditions on that day, screening effect and other possible ground effects were not considered. In the mapping party, a total of 12 users participated in the noise collection. This activity was carried out on Friday 25th October 2013, between 9am and 11am. The devices used on this experiment are shown in Table 1.

Table 1: Devices used to map noise and number of observations taken

Device model	Num. of devices	Num. of samples taken
LG Nexus 4	4	282
HTC One	1	35
HTC Wildfire S	1	67
Samsung Galaxy S4	1	31
Samsung Galaxy S3	1	112
Samsung Galaxy S2	1	1*
Samsung Galaxy Ace 2	1	29
Sony Xperia S	1	21
Celkon A27	1	3*

As seen, there was a reasonable variety of mobile devices. In total, 581 noise samples were taken during the mapping party. Two of the users (marked with an asterisk) had problems with their devices and could not get enough data during the established time. The study area used in this experiment may be considered a small-sized real scenario, where there are multiple devices providing a different number of observations, such as in a real scenario.

Professional data. Every four years, a private company carries out a noise pollution study [14] in Campus following ISO 1996 standard for acoustic reports. In 2012, this study was done for the third time, using the same methodology: Noise measurements are collected during the daytime, when most of human activities occur, and span several days. Campus was monitored following a grid pattern to assure uniform data distribution. In each node, the sound level meter was exposed to noise pollution for 5 minutes. This professional way of data collection considers noise weakening conditions, such as screening effect, sampling height, wind speed or distance to buildings and prevents the noise acquisition of those unwanted conditions.

3.2 Analysis and results

In general, the process of analysis is carried out as follows: Represent the point features on ArcMap 10.2, create an interpolation surface with the Spatial Analyst extension and then subtract the VGI data surface to the professional data surface. For each point, the mean noise and the standard mean error was obtained and those results are presented in a chart. Considering that four participants had the same mobile device, results will be presented in a double way: together for all measurements taken and then for two specific models.

For all noise maps, we chose a color ramp from green to red. Although in general it is useful to detect noisier areas, it is important to consider this when comparing images, because the legend will be slightly different.

General noise map. Using ArcMap 10.2, collected data were represented as a point feature layer. In order to see if there is a spatial relation among the values represented by each point feature, an interpolation surface using Kriging was created. Then, the same operation was repeated with data collected by professional means, so we obtained two basic different maps and compare them. Finally, both raster layers were subtracted using Raster Calculator, to obtain another map showing the difference in measurements of both layers.

Figure 1 depicts the first attempt of creating the campus Noise Map. It is representing the total bulk of data, without any filter correcting possible outliers. As seen, there is a clear similarity between two maps, detecting low noise levels (~50dB) in the central garden and surroundings, moderate noise levels (55dB to 60dB) around the faculties and high noise levels near the road used as main access to campus (>60dB). As seen, most of the samples collected with crowdsourcing present a certain clustering near the nodes of the professional grid.

Figure 2 presents the difference between professional and volunteer raster layers seen in Figure 1. Pink areas highlight places where VGI noise layer measured higher values in decibels while green areas represent the opposite phenomenon.

Finally, areas in yellow areas represent areas where the measurements taken with both methods are very similar. As seen, differences are visually remarkable, but examining the map legend, they are ranging from -5.2dB to 5.6dB, results in line with the ones obtained in [12].

Figure 1: General overview of professional (top) and VGI data (bottom)

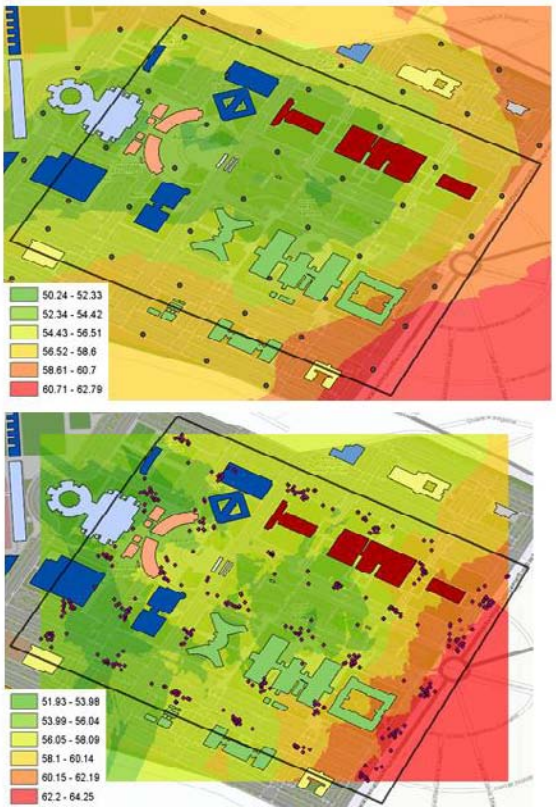
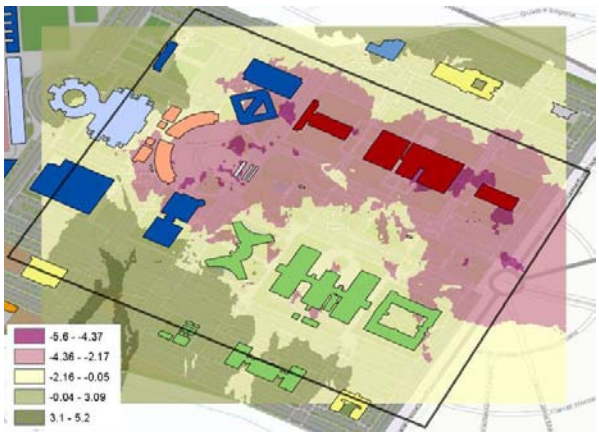
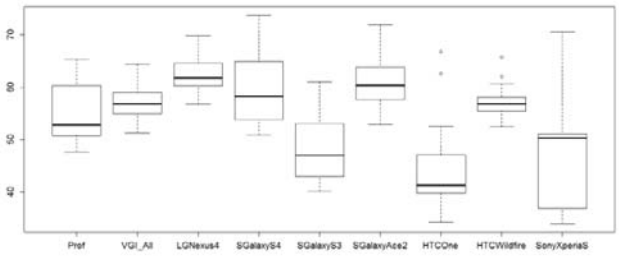


Figure 2: Difference in decibels between professional and raster layers



To conclude this section, a chart (Figure 3) summarizing all data collected in this experiment is presented. First column represents the range of professional noise measurements and the second one is a summary of the subsequent columns. Column "VGI_All" was created by obtaining the mean of the volunteer observations around each node of the grid. It is possible to see that each column in itself is not very similar to the "Prof" one. However, when all individual data are summarized using the mean of all observations per grid node, it seems that results are much more similar to the professional dataset.

Figure 3: Chart showing the summary of professional and volunteer observations collected.



Noise maps for particular mobile devices. In this section new noise maps are created for two particular models: LG Nexus 4 and Samsung Galaxy S4. We chose the first model because four of the devices participating in the experiment were made by this manufacturer and provided one third of the samples. Regarding the second model, we chose it due to its (at present) high-end hardware capabilities. To create these noise maps we carried out a similar process as described in the previous section.

Figure 4: Campus noise map and difference from professional data for LG Nexus 4

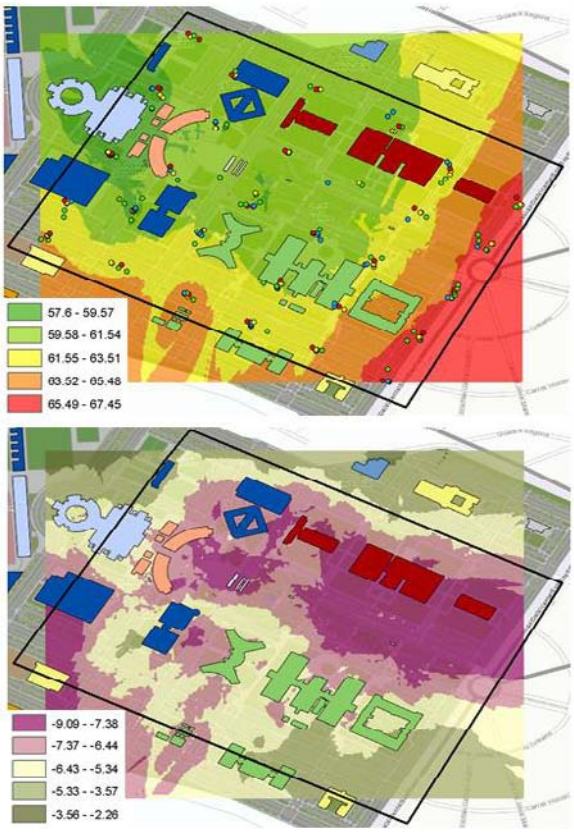
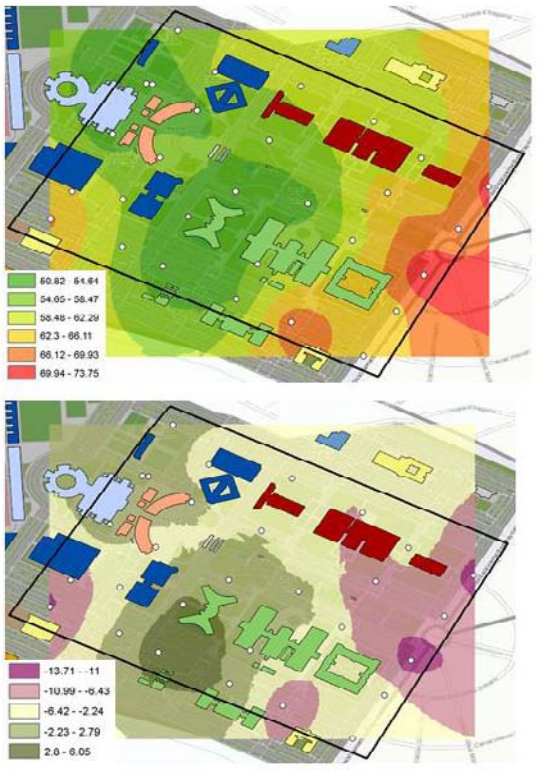


Figure 4(a) presents the general noise map for LG Nexus 4. There are four colored dots in the map showing where each of the four users took a measurement. In general, the map roughly presents an appearance similar to the one obtained with professional data and it is highlighting the same pattern: the campus is quiet in its central part and noise is increasing as long as we get closer to the main roads and accesses. However, when studying the legend, we can see that in this case, the minimum and maximum boundaries are shifted about 5dB higher than the professional samples. Figure 4(b) depicts the difference between both datasets. As seen, there is a large area in dark pink indicating that LG Nexus 4 mobile devices measured between ~7dB-9dB more than the professional dataset. Similarly, Figure 5(a) presents the general noise map for Samsung Galaxy S4. Resulting map is quite different to Figure 1(a) and the device is collecting a wide range of decibels. In this case, the difference layer depicted in Figure 5(b) shows this device is measuring higher mismatches nearby the main access road (from ~6dB to 14dB).

Figure 5: Campus noise map and difference from professional data for Samsung Galaxy 4



Summarizing, it seems that noise maps created for a particular mobile device do no present very accurate results when compared to the professional noise map with the current number of samples. However when all of them are combined into a single one, the output layer presents a reasonable similarity with the professional one.

4 Limitations

The experiment performed in this paper show comparing volunteer data taken with a single device to professional samples does not provide accurate results. However, acceptable noise maps are obtained with the combination of observations provided by a heterogeneous set of devices. Moreover the costs of the required platform (basically maintaining an Internet server) are low with respect to professional tests and the availability of mapping is total. With our proposed system, anyone can take a noise sample in anyplace without time restrictions.

Noise is highly volatile, so, in principle, each sample taken in a determined timestamp might be valid. Several authors [13, 3, 15] point out that it is recommended to take noise samples lasting several minutes to minimize the effect of sudden noises sources. In this experiment, the sampling time lasted several seconds and probably is not enough to provide highly accurate results. Similarly, no sources of attenuation were considered, such as, geometric spreading of noise, physical barriers and we did not use any noise propagation model.

5 Future work

This experiment was carried out using just Android devices in order to obtain a first assessment of noise capture with non-professional means. It would be interesting to repeat this experiment using other devices, such as the ones based on IOS and Windows Phone. Moreover, we also consider using open hardware platforms, such as Arduino or Raspberry Pi, with specific sound level sensors to build a low-cost noise monitoring station.

6 Conclusions

This paper describes a way of comparing volunteer and professional noise data. The professional data was provided by a private company, whereas we generated the volunteer data by means of Android-based devices. Using ArcGIS Spatial Analyst, we created two noise maps from the point features representing noise with a Kriging function. Our results show that individual measures do not seem very reliable, but acceptable results appear when we combine the maps obtained with the different devices used in the experiments.

In general, considering the noise ranges acquired with the professional sound level meter (50dB to 63dB) and the volunteer ones (52dB to 65dB), we consider that noise monitoring through mobile devices is showing very promising results.

We are conscious that this is a preliminary analysis to give a general overview of the potential of VGI data to measure noise pollution. We are not stating that official noise maps and acoustic studies are not needed anymore. However, VGI data might be sufficient for multiple daily situations, like measuring the noise levels on a leisure area (children playground, city center) or for early detection of city noise issues, such as heavy traffic on a residential street. Crowdsourcing noise pollution is a low cost approach that

might be suitable for those communities with a lack of noise sensor networks.

References

[1] J. R. Barber, K. R. Crooks and K. M. Fristrup. *The costs of chronic noise exposure for terrestrial organisms.* Trends in Ecology & Evolution 25(3):180-189. doi: 10.1016/j.tree.2009.08.002, 2009.

[2] M. Craglia. *Volunteered Geographic Information and Spatial Data Infrastructures: when do parallel lines converge?* Proceedings of the Specialist Meeting on Volunteer Geographic Information, Santa Barbara, 2007.

[3] L. Filipponi, S. Santini and A. Vitaletti. *Data collection in Wireless Sensor Networks for Noise Pollution Monitoring.* Proceedings of the 4th IEEE international conference on Distributed Computing in Sensor Systems. Santorini Island, 11-14 June 2008.

[4] C. L. Francis, C. P. Ortega and A. Cruz. *Noise pollution changes avian communities and species interactions.* Current biology 19:1415:1419. doi: 10.1016/j.cub.2009.06.052, 2009

[5] M. F. Goodchild. *Citizens as sensors: the world of volunteered geography.* Available via http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Goodchild_VGI2007.pdf, 2007.

[6] I. Garcia-Martí, L. E. Rodríguez-Pupo, L. Díaz and J. Huerta. *NoiseBattle: A gamified application for Environmental Noise Monitoring in Urban Areas.* Proceedings of the 16th AGILE Conference on Geographic Information Science, Leuven, 14-17 May 2013.

[7] I. Garcia-Martí et al. Mobile Application for Noise Pollution Monitoring through Gamification Techniques. In: Herrlich M, Malaka R, Masuch M (eds) MoGa'12: 4th Workshop on Mobile Gaming. 11th International Conference on Entertainment Computing, Bremen, September 2012. Lecture notes in computer Science, vol 7522. Springer, Heidelberg, p 562, 2012.

[8] M. Haklay and P. Weber P. *OpenStreetMap: User-generated street maps.* IEEE Pervasive Computing 7(4):12-18 2008.

[9] M. Hodson and S. Marvin. *Urbanism in the anthropocene: ecological urbanism or premium ecological enclaves?* City 14(3):299-313. doi: 10.1080/13604813.2010.482277, 2010.

[10] E. Kanjo. *NoiseSPY: A Real-Time Mobile Phone Platform for Urban Noise Monitoring and Mapping.* Mobile Networks and Applications Journal 4(15):562-574.

[11] N. Maisonneuve, M. Stevens and B. Ochab. *Participatory noise pollution monitoring using mobile phones.* Information Polity 15(2):51-71, 2010.

[12] N. Maisonneuve, M. Stevens and L. Steels. *Measure and map noise pollution with your mobile phone*. Proceedings of the DIY for CHI workshop, 27th Annual CHI Conference on Human Factors in Computing Systems, Boston, 4-9 April 2009.

[13] K. R. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu and W. Hu. *Ear-phone: and end-to-end participatory urban noise mapping system.* Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks, Stockholm, 12-16 April 2010.

[14] Re-Ma Ingeniería S.L.*Evaluación del ambiente sonoro en el Campus Riu Sec, Universitat Jaume I de Castelló* (An assessment of noise levels in Riu Sec Campus, University Jaume I of Castelló), 2012.

[15] S. Santini, B. Ostermaier and R. Adelmann. *On the use of sensor nodes and mobile phones for the assessment of noise pollution levels in urban environments.* Proceedings of the 6th international Conference on Networked Sensing Systems, Pittsburgh, 17-19 June 2009.

[16] J. P. Swaddle and L. C. *High levels of environmental noise erode pair preferences in zebra finches: implications for noise pollution.* Animal behavior 74(3):363-368. doi: 10.1016/j.anbehav.2007.01.004, 2007.

[17] M. Swilling. *Reconceptualising urbanism, ecology and networked infrastructures.* Social Dynamics 37(1):78-95. doi: 10.1080/02533952.2011.569997 , 2011.

[18] United Nations (2011) State of the World's Cities 2010/2011 – Cities for All: Bridging the Urban Divide. Accessed 03 Dec 2013. http://www.unhabitat.org/pmss/getElectronicVersion.aspx?nr=2917&alt=1

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Crowdsourced-based mapping of historical west-to-east routes from the textual accounts of European Travelers

Mehdi Ebadi
Institute of Geography
Heidelberg University
Berliner str. 48, 69120
Heidelberg, Germany
mehdi.ebadi@geog.uni-heidelberg.de

Jamal Jokar Arsanjani
GIScience research group
Heidelberg University
Berliner str. 48, 69120
Heidelberg, Germany
jokar.arsanjani@geog.uni-heidelberg.de

Mohamed Bakillah
GIScience research group
Heidelberg University
Berliner str. 48, 69120
Heidelberg, Germany
mohamed.bakillah@geog.uni-heidelberg.de

Abstract

Through the centuries, numerous travellers and orientalists visited Persia (Iran) and described the country and its inhabitants in their travel writings. These travel accounts comprise valuable historical information about the people and their traditions. A literature on travel writings indicate that surprisingly, despite the importance of these recordings, the studies related to the different aspects of these travels, such as the travel routes and the varieties of the possible application of them on the modern time are relatively scarce. The current research deals with the travel routes of nine the most famous early modern European explorers. Accordingly, in addition to digitalizing and mapping the taken routes, the dynamics of their itineraries are analysed.

1 Introduction

Historically, Persia, the land between Inner Asia and Arabian, India and Mesopotamia, from early history has attracted the attentions of several travelers, who visited the country with a variety of motivations such as explorers, diplomats, merchants, missioners, pilgrims, soldiers, etc [1,2]. Through the centuries, numerous travelers and orientalist have described and reported their journeys through different geographical places in their travel writings [2,3]. To name the most famous ones, Herodotus (484-425 BC) was the first known traveler to Persia and Marco Polo (1254 - 1324) after his father Niccolo in a trade expedition to China crossed Persia, and Sven Hedin - (1865-1952) traveled across the two largest Iranian deserts of *Lut* and *Dasht-e Kavir* [1,3,4]. They have recorded remarkable and unique information about the details of their journeys beginning from their origins towards their destinations into a number of documented accounts. Apart from their literary importance, these travel accounts comprise valuable historical information about the chosen routes, the visited places and people, and the existing traditions [1,5]. Furthermore, they have depicted their geographical observations very carefully so that a number of books from their observations and experiences have been published [1-4]. These literature provide us a variety of geographical evidences about the morphology of routes, urban areas, and natural features such as lakes, rivers, deserts, mountains, and so on. These travelogues in fact, offer a kaleidoscopic panorama of a specific region, which is in this case Iran, and permit us to convert their descriptions to map object features. In fact, historical travel writings make the picture of the past come to life by mapping them through the textual evidences in the early modern centuries [7,8]. Considering the periods that European travelers have visited and reported from Persia, the early-modern time might be assumed as the Golden Time. Especially when Shah ‘Abbas I (r. 1588-1629) came to power and ran an outward looking agenda, the country has opened its doors to travelers, and in particular, the Europeans. It has been mentioned in their travel writings that the people in Europe and Iran became familiar with each side’s traditions. The descriptions of the travel routes in the travelogues reveal a lot of information about the history and the historical development of the roads and travel corridors through/in Iran [3,4]. For instance, one can identify the common routes between cities which they were being taken in their itineraries. The travelogues provide us information about the conditions of the travel’s amenities including security, travel permission, overnight and restaurant facilities, etc. The on-the-way cities and their physical appearance as shown in Figure 1 can be rebuilt through mapping their textual reports.

Figure 1: The painted landscapes of Saba city in ancient Persia by Adam Olearius in 1633

The main objective of this study is to crowdsource the travel routes of some travelers based on their explanation of the itineraries in order to map these routes. The most famous European explorers in the early-modern time were chosen for this study, namely Pietro Della Valle (1586-1652), Heinrich von Poser (1599- 1661), Adam Olearius (1603-1671), Sir Thomas Herbert (1606-1682), Jean de Thévenot (1633-1667), and Sir Jean Chardin (1643-1713), Engelbert Kaempfer (1651-1716), Moritz von Kotzebue (1789-1861), and Heinrich Brugsch (1827-1894). To achieve this, their documented itineraries were manually read and the textual accounts were interpreted to delineate their chosen routes. Considering the history of travels of European indicates that from ancient time, Persia was very much part of this general trend and was a coveted destination for European travelers in search of adventure or trade.

2 Materials and methods

From the perspective of tourism studies, the historical routes that were taken by these travelers, not only were one of the important elements and facilitators of these travels, but also are the cultural and historical evidences of these movements. Accordingly, these routes together with the attractive descriptions about tangible and intangible cultural heritages along them ("tourist icons") are assumed as the most important "cultural tourism routes". The textual accounts of the nine abovementioned traveleres were studied and intepreated. The name of known places, regions, and routes were then marked and digitalized on a map.

3 Results

Having the travelers' accounts enabled us to result in a map, which illustrates the approximate geometrical representation of the taken routes. Figure 2 displays them in different symbols.

Figure 2: the primary routes of the nine European Explorers

Visual analysis of the taken routes reveals that the Iran plateau and its southern marine territories were the main corridor of explorations and travels from west to east. To be more precise, while these travellers started their journeys from different places of Europe and they navigated themselves through different corridors until the Iran plateau, they all pursued their routes either through the central Iran or the Persian Gulf sailing ways. This could be due to geographical conditions or cultural attractiveness. The literature review confirms that the latter is the case.

4 Conclusion

In this study, it was intended to highlight the power and potential of crowdsourcing in tourism industry and historical cultural mapping [9-11]. The market of the tourism industry is undergoing widespread and penetrating changes in both the behavioral and technological attributes of the global tourist. The growth in global media and communications are creating an experienced, value-conscious tourist looking for a meaningful interaction with the local communities. Many people no longer want the ''sand, sun, and sea'' of the past [5], but an experiential, multi-activity tourism. Among the significant methods of motivating the tourists to visit a destination (especially in the less-developed regions) is to develop the so-called ''cultural tourism routes''. According to the primary results of the current research, the routes of the pioneer European explorers ''potentially'' have the capacity to be developed as new cultural tourism routes in Iran. Consequently, as future work, the possibility of designing, modelling, and developing these historical routes as cultural and/or literary ''tourism routes'' should be considered.

Through the centuries, travel has been considered as a cultural mean, which has connected the countries and regions. The analysis of cultural routes, as a development and environmental improvement instrument, is undoubtedly among the most interesting topics within the specific scientific community and, perfectly in line with the concept of cultural heritage expressed both on a national and international level within such organizations as INCOMOS, UNESCO, Council of Europe and European Commission. In fact, tourism trails

and routes, offer a means to interpret history, culture and nature [4].

Future works should be directed towards the following specific research objectives, namely to determine:

- The historical aspects of the visits of these European explores in Persia such as the situations of travel enmities (i.e. travel security, the quality of the roads, means of transportation, accommodation and catering facilities etc.) as well as social status, *profession*, *destination*, and motivations of these travelers,
- Digitalizing the travel routes of these European explorers in Persia with more literature and through a semantic approach,
- The tourism ‘‘icons’’ as well as tangible and intangible cultural heritages along these routes should be recognized.

5 References

[1] J. Andersen. *Orientalische Reise-Beschreibungen*. Schleswig 1669.

[2] A. Olearius. *Die erste deutsche Expedition nach Persien, (1635 - 1639)*, Brockhaus, Leipzig, 1927.

[3] F. Ratzel: *Poser, Heinrich von.* In: *Allgemeine Deutsche Biographie* (ADB). Vol. 26, Pages 456 – 458. Duncker & Humblot, Leipzig, 1888.

[4] F. Hayes, D. and N. MacLeod. Packaging places: designing heritage trails using an experience economy perspective to maximize visitor engagement. *Journal of Vacation Marketing*, Vol. 13: 1. Pages 45-58. 2007.

[5] J. Mograbi. *The local economic impacts of tourism in Sodwana Bay*, Unpublished BSc Honours project, school of Geography, Archaelogy & Environmental studies, University of the Witwatersrand, Johannesburg. 2005.

[6] G. Evans. & J. Foord. (2008). Cultural mapping and sustainable communities: planning for the arts revisited. Cultural Trends, 17(2), 65–96. doi:10.1080/09548960802090634.

[7] S. Pile. & N. Thrift. (Eds.) (2013). Mapping the subject: geographies of cultural transformation. Routledge.

[8] P. Poole. (2003). Cultural mapping and indigenous peoples. A report for UNESCO.

[9] C. Heipke. (2010). Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing, 65(6), 550-557.

[10] J. Oomen. & L. Aroyo. (2011). Crowdsourcing in the cultural heritage domain: opportunities and challenges. In Proceedings of the 5th International Conference on Communities and Technologies (pp. 138-149). ACM.

[11] M. Dodge. & R. Kitchin. (2013). Crowdsourced cartography: mapping experience and knowledge. Environment and Planning A, 45(1), 19-36.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Using Open Street Maps data and tools for indoor mapping in a Smart City scenario

Guillermo Amat
FHC25
Roger de Lauria 19 5-B
Valencia, Spain
guillermo.amat@glass.u-tad.com

Javier Fernandez
FHC25
Calle Rozabella, 4
Las Rozas, Madrid, España
javier.fernandez@glass.u-tad.com

Alvaro Arranz
FHC25
Calle Rozabella, 4
Las Rozas, Madrid, España
alvaro.arranz@glass.u-tad.com

Angel Ramos
FHC25
Calle Rozabella, 4
Las Rozas, Madrid, España
angel.ramos@glass.u-tad.com

Abstract

This paper explains the experience of implementing a Smart City scenario using Open Street Maps tools and data. An indoor mapping system including not only a localization and navigation solution, but also a natural speaking environment as a human to machine interface is proposed. The solution is based on a NoSQL database for storing GIS data, a public web service layer used to obtain information, user's current position, navigation routes and human language interaction. An Android mobile client application is used for providing the proper access to all these services. As a case study, the system was successfully implemented in the U-TAD University.

The results shown in this paper can be considered as a demonstration of the previous work related to indoor data representation (IndoorOSM draft) and the navigation solution designed at the Universidade do Minho based on Open Trip Planner. In addition, FHC25 includes a tagging proposal for human language recognition systems.

Keywords: OSM, GIS, Smart Cities, indoor location, indoor navigation, HCI.

1 Introduction

Cities are becoming more intelligent over the time, producing huge amount of data. Citizens living in Smart Cities must have applications that allow access to their services and data. Having them handy, maybe on our smartphones and in the near future accessible in our own wearable technology, is also a challenge.

FHC25 is leading a Smart City Project called Perception[1]. This research project studies different fields related to Smart Cities, such as indoor location, speech recognition and augmented reality. As a testing prototype of the project, an application called Smart U-TAD was implemented at Las Rozas, Madrid. It was conceived as a Smart City prototype bounded to a smaller space: a university building. The idea was to offer information services adapted to new technologies and mobile devices.

In the last few years, several solutions related with these technologies have appeared. Google, probably the most important worldwide map provider, presented his *Google Indoor Maps* [1], oriented to indoor mapping and localization. It is composed of an online indoor map uploading service, indoor visualization technology and a training mobile application for fine localization. However, their usage is restricted to public buildings and their map uploading service is not automatic.

ESRI is also considered one of the leading companies delivering geographical information. Their indoor technology [2] is a complete bundle offering indoor mapping, 2d and 3D visualization and routing. Finally, worth to mention Microsoft's *Bing Venue Maps* [3] indoor mapping service. However, in comparison with the rest of the technologies aforementioned, it is not mature enough and still under development.

The work herein presented studies the problem of creating an information system capable of serving geographical information, accurate indoor positioning and advanced methods of human to machine interaction. For this purpose, Open Street Maps data and tools were used in conjunction with a NoSQL geospatial database, a Restful web interface and a smartphone application.

This paper is organised as follows. Section 2 describes the functional architecture adopted. In Section 3 the mapping task process is explained. Afterwards, Section 4 introduces MongoDB as a geospatial storage solution while Section 5 describes the adopted web service approach. Lastly, Section 6 shows the Android client.

2 System overview

The proposed system's architecture consists of three different elements: a *data layer*, a set of *public REST web services* and a mobile *client*. In order to guarantee secure communications, an additional intermediate layer between the client and server was included, providing certificate-based encrypted communications. Figure 1 shows the basic structure and the relations between layers.

The *data layer* persistence was relayed to MongoDB. Nowadays, there are many applications and websites based on geolocation that require infrastructure for storing and processing geographic information. MongoDB provides this capability and also geospatial queries. In addition, it supports

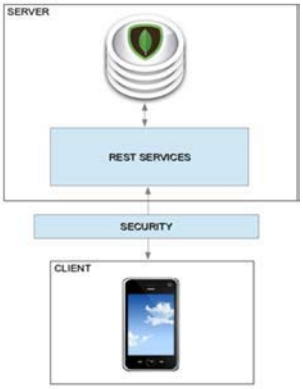
1 This project is framed in the Avanza 2 Plan of the Spanish's Ministry of Industry, Tourism and Trade

REST and JSON specifications and presents high performance and good horizontal scalability [4].

The *data* layer is accessed by the *public REST web services.* REST was used due it is based on some well know standards (HTTP, JSON, URL), is a lightweight protocol compared to SOAP and provides easy scalability.

Finally, the *client* was implemented as an Android mobile application that consumed the REST services and was responsible for the visualization of the indoor maps and additional information such as points of interest in the U-tad University.

Figure 1: Basic system architecture



3 Mapping the building

The university case study was bounded to a single university building. The aim was to provide location services, indoor navigation, information related to the facilities and teachers, etc. This had to be available for students or even the staff which performs various functions in the university.

This section describes the generation and introduction of all the information concerning this building. It was achieved in three steps. First, all spaces and access structures were established following the instructions set out in the OSM's indoor mapping draft [5]. Then, the information was refined adding some more data to automatically generate a series of files needed to model the recognition of spoken language. Finally, we used the example of the Universidade do Minho [6] to create a navigable indoor environment with Open Trip Planner.

There were several requirements such as taking existing data from Open Street Maps, the ability of manipulate easily GIS data, and the capacity to export the results to an XML file or even a JSON format, that made Java Open Street Map editor (JOSM) the best choice for the project. To keep the work safe, the generated files were stored in a Git repository in OSM format. At the moment of needing to export information to our system, the format used was GeoJSON due to it is more appropriate for the treatment and storage in MongoDB.

3.1 Introducing buiding data

The scenario was modeled starting from some images of each of the levels. Following the recommendations stated by – OSM wiki [5, 7], these images were overlaid using PicLayer JOSM plugin and, from there, the boundaries between the rooms were laid out and some other existing elements were drawn. An example is shown in Figure 2.

Figure 2: Drawing U-TAD building from an image using PicLayer.

Once the drawings of all levels were finished, each place was labeled as described in the OSM's draft. This means:

Rooms were tagged using
- buildingpart= room
- name=*

Corridors as
- buildingpart=corridor

Stairs were more complex, requiring more tags:
- buildingpart=verticalpassage
- buildingpart:verticalpassage=stairs
- buildingpart:verticalpassage:floorrange = x to y
- level= z
- name=*

Elevators were a very similar case:
- buildingpart=verticalpassage
- buildingpart:verticalpassage=elevator
- buildingpart:verticalpassage:floorrange= x to y
- level= z
- name=*

Finally, every ***door*** was identified using this key and value:
- door=yes

In addition, contours of each one of the levels were made and labeled as *shield*, also following the guidelines of the draft. Moreover, for each of the levels, a relationship that contained all its elements was created. This was a great help as we advanced in modeling step because it allowed us to filter the items that were shown, so we could work considerably more comfortably. The differences showing the editor before and after including the relationships are shown in Figure 3 and Figure 4 respectively.

Figure 3: All levels and their components show at the same time

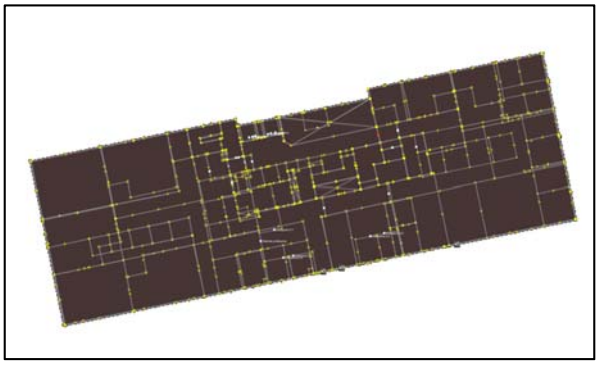
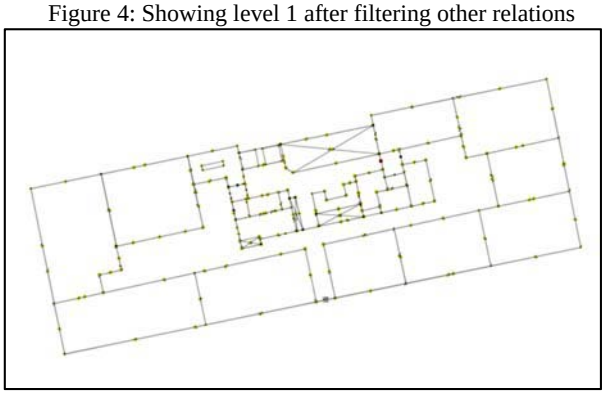


Figure 4: Showing level 1 after filtering other relations



At the end of this first mapping phase, early problems such as the lack of adequate information came to light. For example, there were rooms unnamed or an undifferentiated name. To solve this issue, specific names were assigned to each of these areas, although it is true that they were not official denominations and may therefore be subject to change after a review.

3.2 Tagging for human language recognition

In the developed Smart City application, people can use speech to ask about places by their functionality or by the specific name. The result could be a route or general information regarding the inquired site. Moreover, if the system detects some kind of ambiguity in the request, a dialogue between the user and the device is established in order to determine the concrete thing the user is asking for.

The problem in this area was to identify the correct OSM tags to be assigned with the values of two types of entities defined in the speech recognition service. Essentially, the information to store consists of names and functionalities. Fortunately, OSM specifies *name* and *amenity* tags, which seemed the best option to cover these requirements.

Table 1: Examples of name and amenity tags values.

Name	**Amenity**	
Library	Library	Same values
Reception	Reception	Same values
Office 0 1	Office	
Office 0 2	Office	

However, another point to consider was the existing multi-lingual territories in Spain, and furthermore, the desire to use the solution in other countries, meaning localization problems. As an example of the first case, is easy to find here a Spanish speaker user saying he is a student of “Univerisdad Jaime Primero” (in Spanish) or using “Universitat Jaume Primer” in a Spanish sentence (the university local name).

To solve this issue, we use the name tag with its language extension [8], just as shown here:

- name: Universitat Jaume I (default name)
- name:es: Universidad Jaime I (Spanish name)
- name:ca: Universitat Jaume I (regional language name)

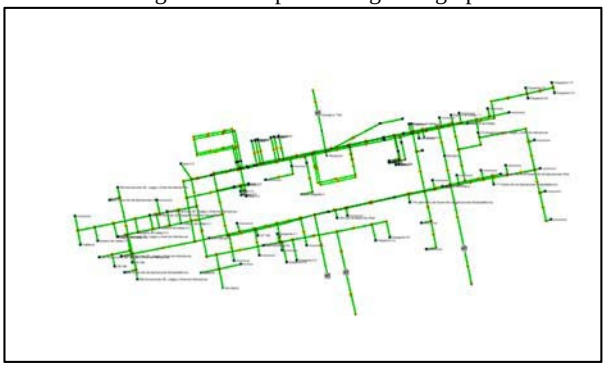
Regarding amenity tag, we found it is being used without accounting for any language restriction. As can be found in Taginfo web site [9], the values are written in any language. In this case there is not a language extension so we used the amenity tag with single-language values and left the translation to a separate process.

After adding all this data in JOSM the results are exported to an OSM file that is batch processed for names and usage extraction, and producing the configuration files needed by the speech recognition system.

3.3 Navigation

This system uses Wi-Fi fingerprinting [14] for positioning. To obtain an indoor navigation system, it was necessary to build a graph representing the ways and possible destination or intermediate points. Then, this information was processed in order to be used with Open Trip Planner [10]. All the subsequent tasks involved were done as established in a previous work [6] from the Universidade do Minho. Again, the tool used was JOSM. The new graph was drawn over the existing maps, so additional filters were set to work properly.

Figure 5: Complete navigation graph



In this representation, the following labels were assigned:

Corridors:

- highway=footway
- indoor=yes
- level=z

Elevators:

- highway=elevator
- name=*

Stairs:

- highway=steps
- indoor=yes
- level=z
- name=*
- oneway=no

Doors:

- door= yes
- level= z
- name=*
- room=yes|class

Building entrances

- door=yes
- entrance=yes

Note that the ways included in this step should be connected to the already existing outdoor roads. This configuration enables route calculation combining both outdoor and indoor spaces.

3.4 Mapping task conclusions

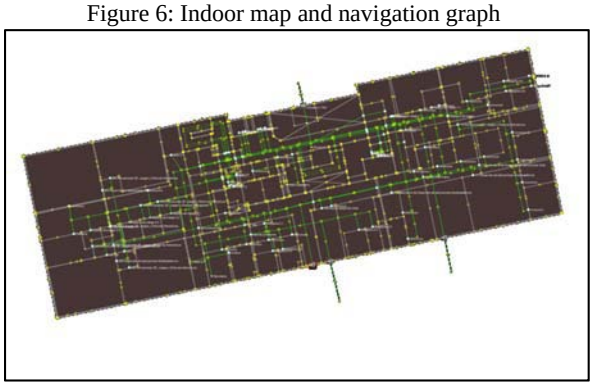
After finishing with these three mapping stages, it was found that the various recommendations are compatible with each other, as they do not contradict the values of the tags suggested in other proposals. What you get at the end is a richer system including all proposed labels.

For instance, regarding doors representation, the comparison in Table 2 can be deduced.

Table 2: Door tagging on different solutions

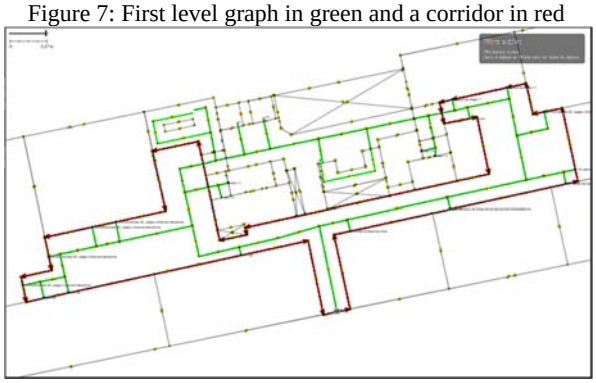
IndoorOSM draft	Universidade do Minho	Percepcion Project
door=*	room=yes level=z name=*	door=* room=* level=z name=*

Figure 6: Indoor map and navigation graph



It also has to be considered the differences between representing a corridor following indoor OSM's draft and drawing the corresponding way for the same corridor in order to navigate. While the first will correspond to a graphic item that will be represented on a map showing the interior of the building, the second will be used as information to calculate routes and will be drawn, in any case, as a segment of a route path. Graphically, these differences are represented in Figure 6 and Figure 7.

Figure 7: First level graph in green and a corridor in red



4 Data layer

The data layer is responsible for storing, among other information, the content of the indoor maps and interest points. The following subsections describe in detail the layer configuration and technologies used.

4.1 Geospatial storage with MongoDB

MongoDB is a NoSQL database oriented to documents. Instead of saving information in tables, as is done in relational databases, this engine saves data structures in JSON type documents with a dynamic scheme (called BSON). Doing that, data integration in certain applications is easier and faster. Geographical data is stored according to the GeoJSON [11] standard. GeoJSON is a collaborative community project that produced a specification for the encoding of this kind of data in JSON format.

Another feature that has affected on a decisive way in its choice for the project is that it is able to support data and geographic queries:

- Proximity queries, where documents are sorted by proximity (nearest to farthest) with reference to a geographical point.
- Bounded queries, whose result is a set of documents that are inside of an area (a rectangle, circle or polygon).

Furthermore, when a query result is returned, it contains the distance to each of the points found.

Regardless of the query type, it is necessary to index the fields storing geographic coordinates with a specific index type called 2dsphere [12]. These indexes support geometries calculation in queries.

Due to these advantages and the performance as a geospatial database, MongoDB is an ideal candidate for projects

requiring storage of geographical data.

4.2 Joining JOSM and MongoDB

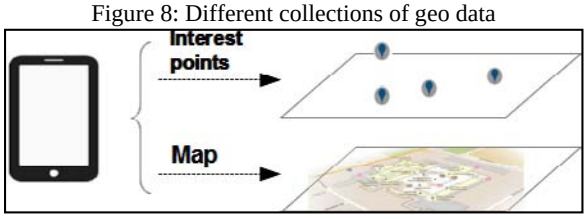
At this point, the system had on one hand a NoSQL engine, which stores points or areas of interest, allowing assign properties to those points or areas and performs searches based on geometry or properties. On the other hand, using JOSM, the data was saved in OSM format (a XML file format that contains all the information needed by Open Street Maps [5]).

In order to convert information between both formats, our system was provided with several processes that transformed everything to GeoJSON format [11]. Such transformation was applied to all the elements of the OSM format: *nodes*, *ways* and *relations* transforming them to GeoJSON geometric elements, which are *points*, *linestrings* and *polygons*.

The XML parsing process is necessary to keep in memory the format and then process and transform it to GeoJSON before storing in the database. After that, a geospatial index is generated in order to optimize queries.

In addition to the map data, we defined a second collection to keep points of interest; this allows separate searches, as if they were different layers, depending on the data that we want to get. To sum up, it is used to save and look up points of interest (places) and geometries at U-TAD University building. Figure 8 represents how both collections are integrated.

Figure 8: Different collections of geo data



Finally, a third collection is used to maintain the routing data to apply indoor navigation capabilities. This means that another process to transform the JOSM painted graph to our internal MongoDB structure had to be developed.

5 Service layer

The project is divided into some sets of services grouped by their functional domain. Any service is offered as a REST web service with their respective resources. Each of them is representing a business concept that can be publicly accessed. The data transport is done via HTTP and the representation is done in JSON format.

5.1 REST services

As said before, all services are divided into four subsets considering its functionality:

1) REST services for creating, updating, deleting and searching points of interest (also called features). These services allow managing different geographical points that want to be highlighted in the application.
2) A REST service for importing a map created in OSM This service transforms a OSM file that contains information about the map to GeoJSON.
3) A REST service for querying maps metadata. This service returns a set of data that client needs to draw the map (maximum longitude and latitude, minimum longitude and latitude, map checksum, etc.)
4) A REST service that queries both features and map areas. It is possible to query different layers because points of interest and map are stored in different collections in MongoDB.
5) REST services for speech recognition and dialogue management. The voice signal is sent to a server where is decoded and processed.
6) REST services providing navigation routes. Departure and destination points are sent to the system and Open Trip Planner calculates the path.

6 Android application

As detailed in the previous sections, the Android mobile application is responsible for displaying all the visual information related to the indoor localization and navigation. This includes displaying the building's surrounding areas, the indoor structure divided into different floors, point of interest (POIs) and navigation information such as paths.

Research on this field has been mainly focused on indoor localization algorithms. However, some visualization analyses can also be found in the literature. Indoor visualization differs from outdoor visualization in the necessity of representing different floors in the same geographical space. Therefore, some concepts of outdoor visualization can be extrapolated to indoor visualization, meanwhile others must be reconsidered.

Note that current OSM tile-based servers are not appropriate for indoor representation. Firstly, most online OSM servers have a zoom restriction that does not permit a close enough visualization. While this issue could be solved changing its configuration, an analysis on its memory performance must be considered. Secondly, the tile approach does not allow representing different floors. Actually, uploading indoor information to OSM servers will lead to rendering all floors one over the other. Thus, a specific application is needed.

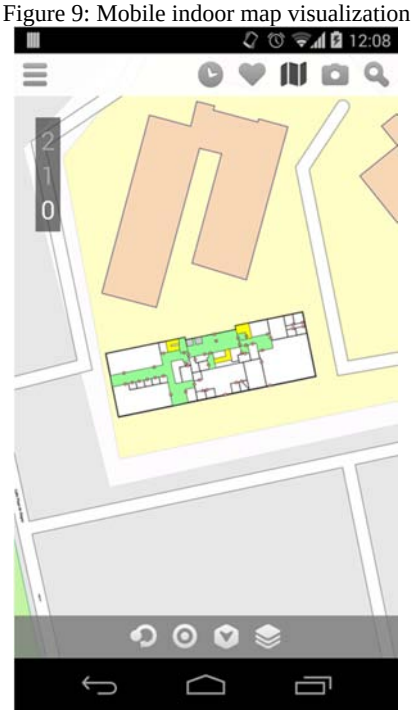
In order to display different building's floors, some authors have taken a 3D approach (Universidade do Minho [6].), while others chose to maintain a planar map representation combined with a selector for navigating between plants (Google Indoors [1], Bing Venue Maps [3], ESRI Indoor [2]). The 3D representation is considered to have some important drawbacks such as the necessity of the three-dimensional modeling of the building, its complexity for intensive use and its understanding difficulties for most users that are not used to 3D model visualization. Thus, a 2D approach was used for the Android application herein presented.

The rendering process is done in the Android device, being comprised of two different steps. Firstly, the background map is rendered. This background has the information equivalent to the one that can be found in the OSM servers, i.e. the

streets, building outlines and orographic information. For this task, the Mapsforge Android library is used. Note that Mapsforge can define different rendering patterns for OSM information [13], so the backgrounds do not necessarily equal the ones found in the servers.

The second step consists of rendering an overlay over the background representing the indoor structure of the building. This indoor information is retrieved using the web-services developed, so they are not hosted in the OSM servers. Finally, a selector is built in the application for switching between different floors. As it is shown in the figure below, different spaces inside the building are represented with different colors. For instance, corridors are represented in green color while elevators or stairways are shown in yellow tones.

Figure 9: Mobile indoor map visualization



7 Conclusions

A system for indoor localization, routing, visualization and map uploading based on Open Street Maps has been proposed. As a case study, the U-TAD University in Madrid was analysed. The system's architecture was divided into three layers (a *data layer*, a set of *public REST web services* and a mobile *client*), allowing easy access to the functionalities.

The system explained in this document extends some other previous research works [5, 6, 7] by including a tagging proposal for speech recognition.

The experience using Open Street Maps data, standards and tools demonstrates an easy and fast deployment of an indoor location solution with minimum cost.

References

[1] Google. *Google Indoor Maps.* Google, 2014. http://maps.google.com/help/maps/indoormaps/

[2] ESRI. *ESRI Indoor Map.* Environmental Systems Research Institute, Inc. http://www.esri.com/industries/logistics/business/gis-indoors

[3] Microsoft. *Bing Venue Maps*. Microsoft Corporation, 2013. http://www.microsoft.com/maps/choose-your-bing-maps-API.aspx

[4] MongoDB Documentation Project. Sharding and *MongoDB.* MongoDB Inc. 2014. http://docs.mongodb.org/v2.4/MongoDB-sharding-guide.pdf

[5] Marcus Goetz, *Indoor Proposal*. OpenStreetMap Wiki, 2014. http://wiki.openstreetmap.org/wiki/IndoorOSM

[6] Nair Isabel Braga Simões Alves, *Uma Soluçao para navegaçao indoor*, bachelor thesis, Universidade do Minho 2012, http://hdl.handle.net/1822/23407

[7] OSM wiki, *OpenStreetMap Indoor Mapping.* OpenStreetMap Wiki, 2014. http://wiki.openstreetmap.org/wiki/Indoor_Mapping

[8] OSM wiki, *Multilingual names*. OpenStreetMap Wiki, 2014. http://wiki.openstreetmap.org/wiki/Multilingual_names

[9] Taginfo, *Search results for amenity tag*. Taginfo Wiki, 2014. http://taginfo.openstreetmap.org/search?q=amenity

[10] Nair Isabel Braga Simões Alves, Jorge Rocha. *OSM indoor:Moving Forward.* Universidade do Minho, 2012. http://ogrs2012.heig-vd.ch/public/ogrs2012/slides/Alves.pdf

[11] Howard Butler, Martin Daly, Allan Doyle,Sean Gillies, Tim Schaub, Christopher Schmidt. *The GeoJSON Format Specification*. GeoJSON, 2008. http://geojson.org/geojson-spec.html

[12] MongoDB Documentation Project. *MongoDB Documentation.* MongoDB Inc. 2013. http://docs.mongodb.org/manual/MongoDB-manual.pdf

[13] Mapsforge Project. *Mapsforge. Free mapping and navigation tools.* https://code.google.com/p/mapsforge/

[14] Nelson Marques, Filipe Meneses and Adriano Moreira. *Combining similarity functions and majority rules for multi-building, multi-floor, WiFi Positioning.* 2012 International Conference on Indoor Positioning and Indoor Navigation

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Comparing Knowledge-Driven and Data-Driven Modeling methods for susceptibility mapping in spatial epidemiology: a case study in Visceral Leishmaniasis

Mohammadreza Rajabi[1], Ali Mansourian[1], Petter Pilesjo[1], Finn Hedefalk[1],Roger Groth, Ahad Bazmani[2]
1: GIS Center, Department of Physical Geography and Ecosystem Science, Lund University, Sweden
Sölvegatan 12, 22362, Lund, Sweden
2: Infectious and Tropical Diseases Research Centre, Tabriz University of Medical Sciences, Tabriz, Iran
{Mohammadreza.Rajabi ,Ali.Mansourian, Finn.Hedefalk,Roger.Groth}@nateko.lu.se,
Petter.Pilesjo@gis.lu.se , bazmany_ahad@yahoo.com

Abstract

The aim of this study is to compare knowledge-driven and data-driven methods for susceptibility mapping in spatial epidemiology. Our comparison focuses on one of the arguably most important requisites in such models, namely predictability. We compare one data-driven modelling method called Radial Basis Functional Link Net (RBFLN - a well-established Neural Network method) with two knowledge-driven modelling methods, Fuzzy AHP_OWA and Fuzzy GIS-based group decision making (multi criteria decision making methods). These methods are compared in the context of a concrete case study, namely the environmental modelling of Visceral Leishmaniasis (VL) for predictive mapping of risky areas. Our results show that, at least in this particular application, RBFLN model offers the best predictive accuracy.

Keywords: Visceral Leishmaniasis (VL), spatial epidemiology, prediction, knowledge-driven method, data-driven method.

1 Introduction

As a major epidemiological hazard, Visceral Leishmaniasis (VL) (commonly known as kala-azar) accounts for a great number of human fatalities, and causes significant damage to public health in developing countries especially poor and rural areas [1, 5, 8, 12, 3]. In order to mitigate losses and damages, many spatial susceptibility studies have been conducted to map the locations that are prone to VL outbreak [1, 4, 7, 14].

Most of the studies about spatial epidemiology assume that disease susceptibility is related to specific predisposing factors and that susceptibility can be assessed as long as the predisposing factors and the relationships between the factors and the disease are identified [1]. The mentioned factors are considered to be the intrinsic nature and condition of the environment, which make the area susceptible to be infected but do not actually trigger an outbreak [12]. In this study, we are comparing three popular methods in the context of VL spatial epidemiology: Radial Basis Functional Link Net (RBFLN), Fuzzy Analytical Hierarchy Process (AHP)-OWA (Ordered Weighted Averaging), and Fuzzy Group decision making. Accordingly, the common predisposing factors for VL are land use/land cover, meteorological factors (rainfall, temperature), topographical factors (altitude, river) and socio-economic factors (access to health-centres, lifestyle) [13]

Knowledge driven and data driven strategies reflect two different perspectives in spatial modelling. More specifically, a knowledge driven approach is based on evidence of varying quality, guidelines, and experts' opinions, while a data driven approach is solely based on the observational data.

This paper presents a comparative approach to disease-susceptibility mapping, which discusses the pros and cons of data-driven approaches versus knowledge-driven approaches. The study is exclusively concerned with VL endemic areas.

2 Materials and methods

2.1 Study area

The study is focused on two districts in Iran including about 800 villages: Kalaybar in the western part of East Azerbaijan province (47.0427° E, 38.864° N), and Ahar, located immediately south of Kalaybar (47.068° E, 38.472° N).

2.2 Data collection

In collaboration with the Infectious and Tropical Diseases Research Centre of the Iranian ministry of health, we collated VL notification data at the village-level, either from central registers or from district centres. Then the information were integrated into one database.

Based on [13], eight items were chosen to be the fundamental factors for predictive mapping of VL risky areas for this research: temperature, precipitation, proximity to rivers, altitude, presence of health-centres, land cover, density of dogs, and presence of nomads

2.3 MCDM

Multi criteria decision analysis (MCDA) is a knowledge-driven transparent process supporting decision-makers faced with making numerous, sometimes conflicting, evaluations by highlighting these conflicts aiming to find a compromise. GIS–MCDA is a process that combines geographical data (map criteria) and value judgments (decision-maker preferences and uncertainties) to obtain appropriate and useful supporting documentation [9].

2.3.1 Fuzzy AHP_OWA

Fuzzy AHP_OWA is a knowledge-driven method in which the degree of risk and trade-off of decision making can be modelled properly. In this approach, we accomplished the two first steps of the AHP at the first stage. In this regard,, the hierarchical structure of the model would be formed, and the relative importance of the predisposing factors would be determined by conducting pairwise comparisons. At this point, the quantifier-guided OWA methods take the lead for the rest of the analysis. The procedure at this stage involves three main steps [10]: (i) identifying the linguistic quantifier Q, (ii) generating a set of ordered weights associated with Q, and (iii) computing the overall evaluation for each i^{th} location (alternative) at each level of the hierarchy by means of the OWA combination function.

2.3.2 Group Decision Making

Group decision-making is a situation in which individuals cooperatively make a choice from the existing options. Applying GIS–MCDA for group decision-making forms aggregated individual judgments into a group preference in a manner in which the best compromise can be recognized [2]. Although the GIS–MCDA approaches have traditionally focused on the MCDA algorithms for individual decision-making, significant efforts have been made to integrate spatial epidemiology for group decision-making settings.

A fuzzy majority approach has been introduced [11] to model the concept of majority opinion in group decision-making problems. Using a linguistic quantifier, the fuzzy majority concept can generate a group solution that corresponds to the majority of the decision-makers' preferences. The linguistic quantifier leads the aggregation process of the individual judgments in such a way that there is no need for rankings of the alternatives of individual solutions.

2.3.3 RADIAL BASIS FUNCTIONAL LINK NETS

The purpose of an Artificial Neural Network (ANN) is to build a model of data-generating process through a learning algorithm. ANNs generally consist of several neurons, which are organized in three layers: input, hidden and output. Looney [6] introduced a modified architecture of ANN termed radial basis functional link nets (RBFLN). The main difference of RBFLN is the use of additional links between the input layer and output layer. These extra lines and weights model the linear part of the input–output transformation [6].

The RBFLN network requires two sets of training points: one that defines the presence of the objects or conditions to be predicted (i.e., VL endemic areas) and a second that defines the absence of these objects (i.e., locations where VL incidence are known not to be endemic). The two sets of points are combined as training data.

3 Results and discussion

In the first knowledge-driven approach the specified environmental factors were first entered into a fuzzy AHP-OWA algorithm to identify susceptive areas in relation to the prevalence of VL. At the first stage, based on the experts opinions (who are local medics and VL specialists) the factors were then classified as "climate" and "intensity of contagion" classes. Temperature, precipitation, rivers, altitude and land cover factors were considered to belong to the "climate" class. The impact of health centres, nomads and density of dogs was assigned to the "intensity of contagion" class.

In the next stage, after structuring the criteria, a pair-wise comparison between factor maps was performed according to their effects on VL. The process was indirectly dependent on the knowledge of experts. By weighting of the AHP, the relative importance of each criterion was obtained. For example, in the "climate" class, the weights that were achieved by AHP were as follows: altitude = 0.45, precipitation = 0.263, distance to river=0.103 and temperature = 0.155. Considering the coefficient Consistency Ratio (CR) = 0.015, i.e. < 0.1, the weight values were validated and remained in the calculations. Figure 1b, shows the result map from AHP_OWA.

Effective factors and parameters associated with VL outbreaks have been entered in the prediction models (even where VL was epidemic).

In the AHP_OWA approach, the achieved prediction data and the registered cases of VL in infected areas have been compared together. When relating risk maps with the infected villages and available information about the patients, the output map indicated that all of the current highly infected villages were predicted to be hazardous areas by Fuzzy AHP-OWA (Figure 1b).

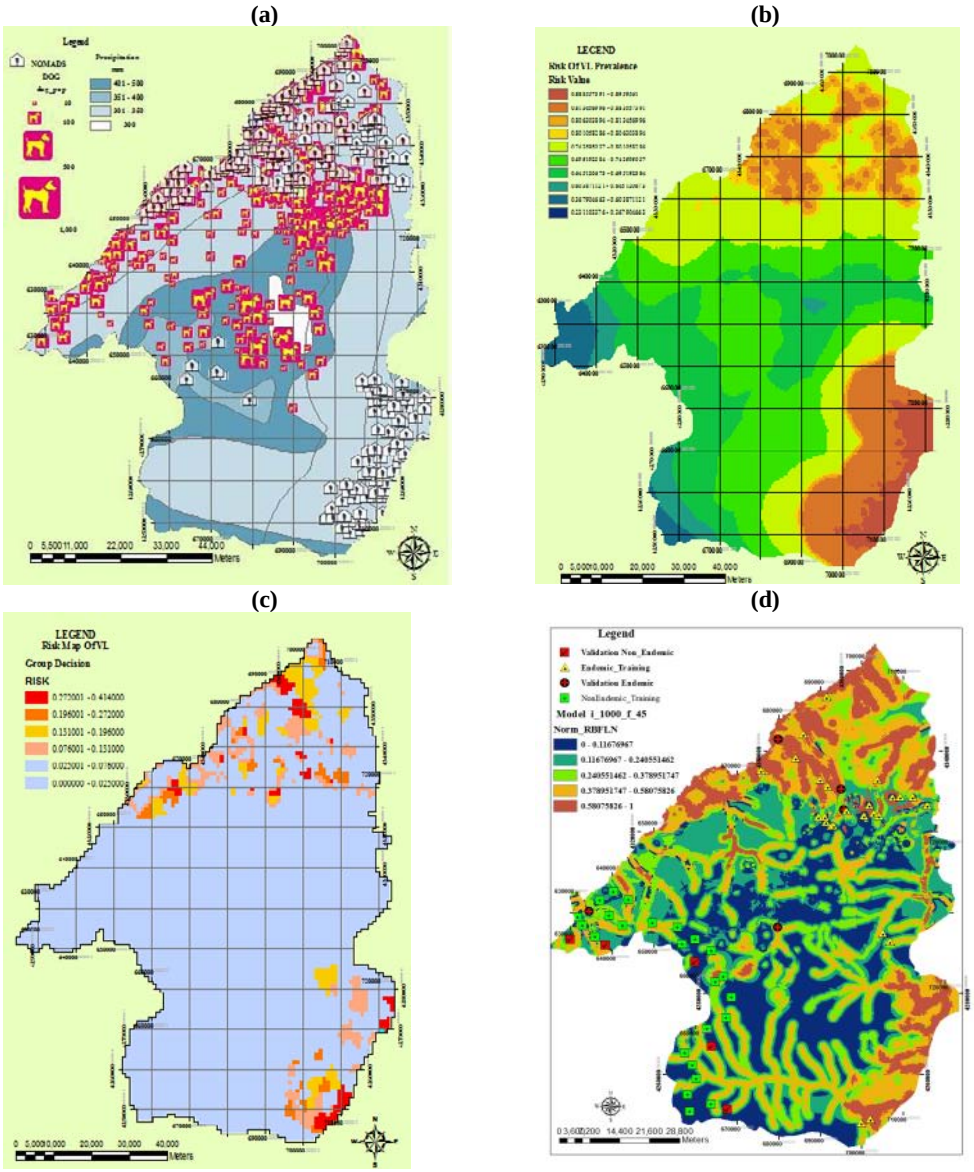
Then the knowledge of five local experts in the field of VL was generalized in a fuzzy group decision-making process. The main objective was to investigate the current situation of the villages at risk to provide urgent emergency services (Table 1).

Table 1: The Opinions of five local VL specialist about the degree of effect of eight VL parameters

Experts	Nomad	Height	Temp	Rain	River	Health Centres	Dog
1	VH	H	M	H	M	M	VH
2	H	M	M	M	H	M	VH
3	VH	VH	H	H	H	H	M
4	H	H	M	L	M	VH	H
5	H	M	M	M	H	H	VH

VH = Very high , H=High , M=Medium , L=Low , VL=Very Low

Figure 1: (a) Distribution of nomadic villages (b) Output map of Fuzzy AHP_OWA (c) Output map of group decision making (d) Output map of RBFLN.



After gathering information and opinions of five local experts about VL and weighting factors by converting the fuzzy terms to hard numbers, the information was combined at various levels of risk and trade-off using fuzzy linguistic quantifiers (Table.1). On the basis of the knowledge of each of the experts, one thematic map was generated. In each of the generated maps, different levels of risk were assigned to the villages (Figure 1). There should therefore be a fuzzy group decision-making process to identify the villages in which most of local experts and medics agree about the severity of the crisis. The risk level for each area was calculated using a fuzzy majority approach in a fuzzy group decision-making process. A new map was generated that indicates the level of danger for each village. The new map should be useful for prioritizing the provision of the health measures for each village (Figure 1b).

The Carl Looney's RBFLN algorithm that was implemented in Arc Spatial Data Modeller (ArcSDM) has been applied [6]. To generate the input exploratory data for RBFLN in the planned model for VL, the evidential maps were overlaid to create a unique conditions grid. A unique conditions grid consisting of 2699 unique overlay conditions, which is a relatively large number, was generated. In the attribute table of the unique conditions grid, there is one record for each unique overlay condition as well as one field for each evidential map. Thus the unique overlay conditions are n-dimensional (n = number of evidential maps) input vectors. The resulting unique condition grid was the input for the RBFLN.

For the purpose of modelling using RBFLN, first, an optimum structure for RBFLN in terms of the number of hidden functions as well as the number of iterations for RBFLN training had to be determined. An RBFLN structure with 45 hidden functions and 1000 iterations, resulting in a summed-squared error (SSE) equal to 0.00378, was considered as the most proper one.

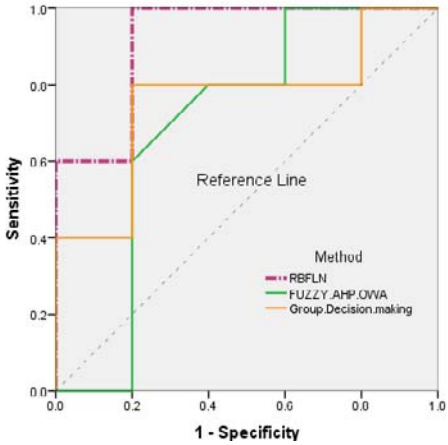
Figure 1d shows the result of applying the RBFLN to create a multiclass predictive map for VL. This map is interpreted as susceptibility of the individual cells in the area in relation with the VL endemicity.

Using the prediction-rate method, the results of the three susceptibility maps were validated by comparing them with the existing infected areas. The prediction rate can explain how well the VL prediction model predicts VL endemicity. In this study, the prediction-rate results were obtained by comparing the infectious villages in the validation dataset with the three VL susceptibility maps.

The areas under the prediction-rate ROC curves (AUC) were calculated. An AUC equals to 1 indicates perfect prediction accuracy (Lee and Dan, 2005).

When ROC curves of these three methods were considered together, their overall performances are seen to be close to each other. The most successful method is the RBFLN model. According to the obtained AUC, RBFLN has slightly higher prediction performance than Fuzzy AHP_OWA and Group Decision Making (Figure 2). This may be due to the fact that in the RBFLN model, the training process makes the data richer, and this enrichment makes the RBFLN slightly more successful than knowledge-based models.

Figure 2: The areas under the prediction-rate ROC curves (AUC)



4 Conclusion

In this study, the application of one data-driven method, RBFLN, and two knowledge-driven methods (Fuzzy AHP_OWA and fuzzy group decision making) has been explored for predictive mapping in spatial epidemiology for VL disease.

The results indicate that, in this particular application, the RBFLN model obtained the best predictive accuracy. Therefore this model may be preferred when mapping the VL susceptibility. Nevertheless, the knowledge-driven methods are also capable of reliably mapping areas of high risk for VL, and they can easier map the risk and trade-off from the decision makers' opinions.

References

[1]. D. S. Barbosa, V. S. Belo, M. E. S. Rangel, and G. L. Werneck, "Spatial analysis for identification of priority areas for surveillance and control in a visceral leishmaniasis endemic area in Brazil," *Acta Tropica*, 131; 56–62, 2014.

[2]. S. Boroushaki and J. Malczewski. Using the fuzzy majority approach for GIS-based multicriteria group decision-making. *Computers & Geosciences*, 36; 302–312, 2010.

[3]. F. Chappuis, S. Sundar, A. Hailu, H. Ghalib, S. Rijal, R. W. Peeling, ... and M. Boelaert. Visceral leishmaniasis: what are the needs for diagnosis, treatment and control. *Nature Reviews Microbiology*, 5(11); 873-882, 2007.

[4]. S. A. Correa Antonialli, T. G. Torres, A. C. Paranhos Filho, and J. E. Tolezano, "Spatial analysis of American Visceral Leishmaniasis in Mato Grosso do Sul State, Central Brazil," *Journal of Infection*, vol. 54(5); 509–514, 2007.

[5]. S. Garg, R. Tripathi, and K. Tripathi, "Oral mucosal involvement in visceral leishmaniasis," *Asian Pacific Journal of Tropical Medicine*, 6(3); 249–250, 2013.

[6]. C. Looney. Radial basis functional link nets and fuzzy reasoning: *Neurocomputing*, 48; 489–509, 2002.

[7]. M. S. Fernández, O. D. Salomón, R. Cavia, A. A. Perez, S. A. Acardi, and J. D. Guccione, "Lutzomyia longipalpis spatial distribution and association with environmental variables in an urban focus of visceral leishmaniasis, Misiones, Argentina," *Acta Tropica*, 114(2); 81–87, 2010.

[8]. T. Hazratian, Y. Rassi, M. A. Oshaghi, M. R. Yaghoobi-Ershadi, E. Fallah, M. R. Shirzadi, and S. Rafizadeh, "Phenology and population dynamics of sand flies in a new focus of visceral leishmaniasis in Eastern Azarbaijan Province, North western of Iran," *Asian Pacific Journal of Tropical Medicine*, 4(8); 604–609, 2011.

[9]. J. Malczewski and C. Rinner. Exploring multicriteria decision strategies in GIS with linguistic quantifiers: A case study of residential quality evaluation. *J Geograph Syst*. 7; 249–268, 2003.

[10]. J. Malczewski, J. Multicriteria decision analysis for collaborative GIS. In: Balram, S., Dragicevic , S. (Eds.), *Collaborative Geographic Information Systems*. Idea Group Publishing, Hershey, 167–185, 2006.

[11]. G. Pasi and R. R. Yager, Modeling the concept of majority opinion in group decision-making. *Information Sciences* 176; 390–414, 2006.

[12]. A. T. Peterson, R. S. Pereira, and V. F. de C. Neves, "Using epidemiological survey data to infer geographic distributions of leishmaniasis vector species," *Revista da Sociedade Brasileira de Medicina Tropical*, 37(1); 10–14, 2004.

[13]. M. Rajabi, A. Mansourian and A. Bazman. Susceptibility mapping of visceral leishmaniasis based on fuzzy modelling and group decision-making methods, *Geospatial Health* 7(1); 37-50, 2012.

[14]. L. Saraiva, J. D. Andrade Filho, A. L. Falcão, D. A. A. de Carvalho, C. M. de Souza, C. R. Freitas, C. R. Gomes Lopes, E. C. Moreno, and M. N. Melo, "Phlebotominae fauna (Diptera: Psychodidae) in an urban district of Belo Horizonte, Brazil, endemic for visceral leishmaniasis: Characterization of favored locations as determined by spatial analysis," *Acta Tropica*, 117(2); 137–145, 2011.

Session:
Visualization

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science. Castellón, June. 3-6. 2014. ISBN: 978-90-816960-4-3

How to visualize the geography of Swiss history

André Bruggmann
Department of Geography
University of Zurich
Winterthurerstr. 190
8057 Zurich, Switzerland
andre.bruggmann@geo.uzh.ch

Sara I. Fabrikant
Department of Geography
University of Zurich
Winterthurerstr. 190
8057 Zurich, Switzerland
sara.fabrikant@geo.uzh.ch

Abstract

Efficient and effective access to and knowledge construction from massively growing spatial and non-spatial databases available online today have become major bottlenecks for the rapidly evolving information society at large. We present a geovisual analytics framework to deal with spatio-temporal knowledge extraction from rapidly growing, and increasingly massive, digital text databases largely untapped for spatio-temporal analyses. Our interdisciplinary, theory-driven approach combines text data mining methods, currently employed in GIScience and geovisual analytics, to re-organize and visualize a semi-structured online dictionary about Swiss history, made available to the general public. We automatically extract spatial, temporal, and thematic information from the text archive, and make it visually available to an information seeker interested in Swiss history, through empirically validated spatialization display techniques (e.g., network visualizations and self-organizing maps). In this case study, we specifically illustrate how spatial relationships between Swiss toponyms can be extracted, analyzed, and visualized using our proposed approach. With this interdisciplinary geovisual analytics approach situated at the nexus of digital humanities, information science, and GIScience we hope to provide new transdisciplinary solutions to facilitate information extraction of and knowledge generation from information buried in vast unstructured text archives.
Keywords: geovisual analytics, geographic information retrieval, information visualization, text mining, digital library.

1 Introduction

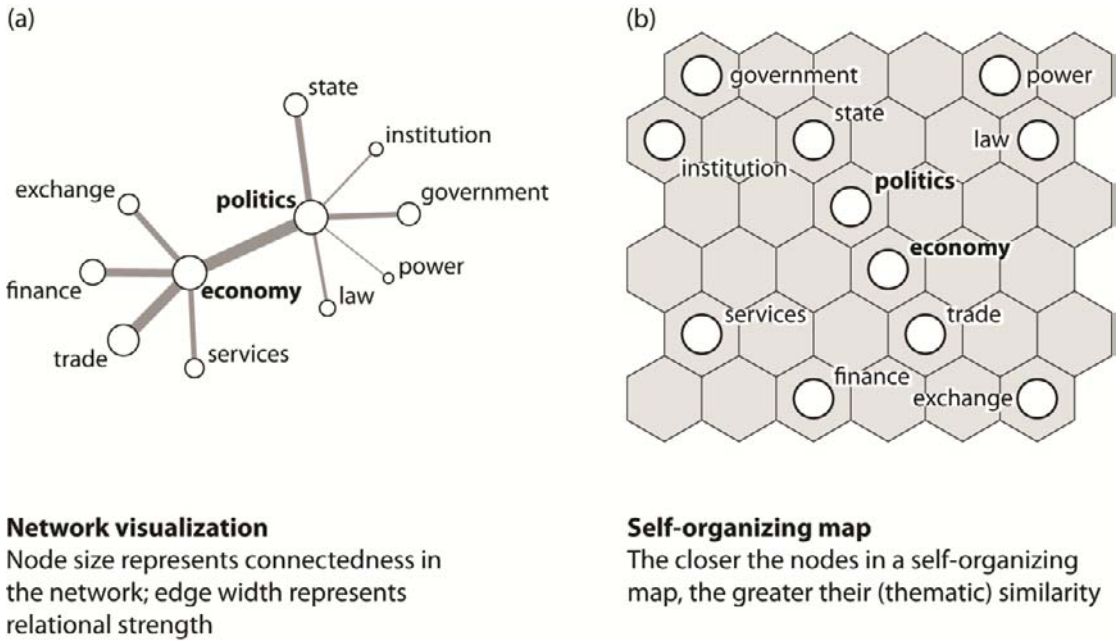
Large online digital libraries such as, the Encyclopedia Britannica or Google Books, for example, provide today's information seekers access to massive collections of unstructured digital text data and diverse multimedia content. Libraries play an important role in the humanities and the social sciences, where text documents have been central data sources for a very long time before digitization, but they are still largely untapped for spatio-temporal analyses. With massive text collections becoming available digitally, knowledge generation challenges have emerged, tackled by a variety of research communities besides GIScience (i.e., computer science, information science, digital humanities, etc.). In this context, automated text analytics techniques and tools provided by the geographic information retrieval (GIR) community coupled with powerful visuo-spatial geovisual analytics (GeoVA) interfaces seem especially relevant. GIR deals with automatically extracting relevant spatio-temporal information from digital text archives across place, over time, and on various topics and themes. GeoVA is concerned with making this information available to an information seeker through powerful, interactive graphic displays to facilitate information exploration and knowledge construction. Both GIR and GeoVA aim at revealing hidden patterns in large, typically unstructured text databases, and in doing so allow users to more easily explore emerging relationships between documents, and eventually increase sense making from vast text databases.

We present an interdisciplinary, theory-driven approach which combines GIR and GeoVA methods to re-organize and visualize semi-structured texts from a digital dictionary about Swiss history. We aim to re-organize spatial, temporal, and thematic characteristics automatically extracted from the text data archive, and make this information visually available to an information seeker interested about Swiss history. Our interdisciplinary approach may provide likeminded GIScientists new inspirations to further advance spatio-temporal methods and tools for similar kinds of datasets in a transdisciplinary context.

2 Background

A variety of automated methods have been developed in GIR to extract spatio-temporal and thematic information from digital text data sources. For example, [8] demonstrate how toponyms (e.g., London) can be automatically extracted from a massive, historical, unstructured text archive, and present novel approaches to deal with disambiguation problems (e.g., does London refer to the Capital of the U.K., or the town in Ontario, Canada?). Other toponym disambiguation solutions can be found in [7], [15] or [17]. Significant advances have also been made in automatically extracting temporal information from digital text sources. [1] and [2], for example, review current research trends in temporal information retrieval. Various date extraction tools (e.g., HeidelTime) have been developed and tested so far [24]. Similarly, attractive solutions to automatically identify thematic relations in text sources, based on the quantification of thematic similarity of texts (i.e., by using word similarity or the similarity of parts of sentences in texts), have been proposed by [23]. Their *Probabilistic Topic Models* (TM) are already widely used in the digital humanities communities. [20]

Figure 1: Network visualization (a) and self-organizing map (b).



demonstrated how the results of the TM approach may be input to graph-theoretic clustering methods, to automatically assign a thematic label (e.g., economy, politics, etc.) to grouped text documents [3].

Moreover, the visualization communities have proposed solutions to depict multivariate (i.e., spatial, temporal and thematic) numerical and non-numerical data. GeoVA has already been successfully employed to visualize uncovered relations in vast text databases. For example, [9], and [22] introduce the *spatialization framework*, which includes a systematic approach to transform high-dimensional data sets into lower-dimensional, spatial representations for facilitating data exploration using spatial metaphors. Self-organizing maps (SOM) and network maps are good examples of this (see Figure 1). Both mapping approaches depict documents that are similar in content close to one another in the visualization, as illustrated in Figure 1. Emerging themes and respective document clusters are schematically represented in Figure 1. The network map (1a) depicts text documents as nodes on a relational semantic network. Document clusters that are similar in content are connected with one another. The bigger the nodes, the stronger the connectedness of the respective document clusters with other documents in the database. The thicker the edge between nodes in the network, the stronger the thematic relationship between two corresponding document clusters in the database.

SOMs (a neural network method) project multivariate input data onto a two dimensional, topological space, typically represented by a regular tessellation (i.e., hexagons) [14]. The neurons in the SOM have the same attributes as the input data. They are placed near each other in the map if they share similar attributes, and are therefore similar in content [21]. The original data are then mapped as points onto neurons with thematically most similar attributes, thus documents of similar content cluster with respective neurons, as illustrated in Figure 1b.

The *spatialization framework* has been applied in various ways for knowledge exploration from vast multivariate (numerical) datasets, including the temporal data dimension. For example [6] applied the SOM techniques to visually explore multivariate quantitative (e.g., census) and qualitative (e.g., open-ended survey responses) data sets. The SOMs facilitated the analysis of the characteristics of survey respondents, the socio-demographic characteristics of San Diego neighborhoods, and the characteristics of the utterances respondents used to describe these neighborhoods.

3 Case Study

Based upon the empirically validated *spatialization framework* by [9] we now outline our interdisciplinary visual text analytics approach applied to the Historical Dictionary of Switzerland (HDS) [12]. We chose this particular dataset for several reasons. As a typical example of an online digital library it specifically contains spatial, temporal, and thematic information. It serves as a proto-typical secondary data resource for researchers and the general public interested in Swiss history. The multi-lingual HDS (i.e., German, French and Italian) consists of 36,188 articles related to the history of Switzerland. For our case study we chose to analyze only the German version of the HDS. The dictionary is structured by article categories, such as *thematic contributions* (e.g., events, institutions, etc.), *geographical entities* (e.g., municipalities,

Cantons, etc.), *biographies*, and by articles about historically important *families*. Currently, the articles are organized in alphabetical order. There are no possibilities to query the articles according to spatial or temporal criteria.

Our visual text analytics approach includes two phases: automated text analytics and visualization. In a first step, we extract spatial, temporal, and thematic information from the HDS articles, using the well-established GIR methods mentioned above (i.e., [8]). We extract historically relevant locations (i.e., toponyms) by first identifying candidate toponyms in the HDS with the Swissnames gazetteer [25]. This gazetteer consists of 156,755 toponyms occurring on Swiss topographic maps on a scale of 1:25,000. Following that, we resolve disambiguation issues, as described in [8]. We then employ *HeidelTime*, a tool developed by [24] to extract historically relevant dates. *HeidelTime* is based on the TIMEX3 annotation standard and the markup language TimeML [18, 19]. *HeidelTime* allows to automatically retrieve dates (e.g., 07/09/2002), periods of time (e.g., 17th century), and other temporal information from texts. Finally, we employ the above-mentioned TM method [23], available in the Text Visualization Toolbox (TVT) in MATLAB [11]. This text analytics phase yields automatically retrieved information, structured in data tables including spatial, temporal, and thematic information.

In the second phase of our approach, we visualize the retrieved information. The main aim is to create an interactive proof-of-concept interface which allows users to gain new insights into the history of Switzerland, based on the re-organization of spatial, temporal, and thematic relations extracted from articles stored alphabetically in the HDS.

In the next sections we illustrate this by example, concentrating on spatial data automatically extracted from the HDS text archive. The employed GIR algorithm extracted 13,719 distinct toponyms (e.g., names of cities, municipalities, villages, historical places, and water bodies), that appear at least once in the HDS database. In total, we retrieved 169,094 toponyms. Table 1 below details the extracted Swissnames toponym categories, the total number of toponym occurrences per category, and their corresponding percentages.

Table 1: Swissnames categories and toponym occurrences.

Toponym Categories	Total	Percent
Cities, municipalities, villages	137,751	81.5
Areas (e.g., forests)	14,693	8.7
Single objects (e.g., churches, castles)	6,917	4.1
Rivers and lakes	4,972	2.9
Mountains	2,152	1.3
Valleys	1,726	1.0
Passes	544	0.3
Miscellaneous	339	0.2
Total	169,094	100

Populated places such as cities, municipalities and villages account for 81.5 percent of all toponym occurrences in the HDS. Given that "people make history", it is not surprising that this kind of spatial information is the most relevant in the HDS. The remaining 18.5% toponym occurrences are non-urban areas (e.g., forests), individual features (e.g., churches, castles, etc.), water bodies, and landforms such as, mountains, valleys, and passes, including other miscellaneous objects.

We then generate a network spatialization, similarly to Figure 1 with the extracted toponyms. Following [10]'s approach, we assume a relationship between two toponyms, if they both co-occur in the same HDS article. The overall strength of a relationship between two toponyms represents the sum of co-occurrences (divided by two) across all articles where both toponyms co-appear at least once. This weighted toponym matrix is then input to the Network Workbench (NWB) tool [16].

4 Results

Figure 2a depicts a spatialized network consisting of toponyms that, overall, occur at least 360 times in the HDS. In other words, toponyms co-appear on average once per every 100 articles. We identified 43 toponyms that meet this requirement. We excluded thirteen toponyms from this set, as they are considered stop words (e.g., *castle*). The GIR algorithm did not already exclude these. The remaining thirty toponyms account for 28.8% of the overall toponym occurrences in the HDS.

We chose the GEM layout algorithm to visualize the network, as to avoid edge crossings. We chose to only visualize the structural most important relationships using a minimum spanning tree (MST) pathfinder algorithm. We ran the Blondel community detection algorithm [3] on the weighted input matrix in NWB, to identify groups of similar toponyms which are depicted as nodes in the network (a), and in the map (b) in Figure 2. The Blondel community algorithm detects toponym groups that have strong within-group relationships, and separates toponyms that only have weak relationships. The detected communities are shown as differently colored nodes in Figure 2. We apply the visual variable line width to depict the strength (i.e., weight) of toponym relationships. The importance of a toponym in the network is shown by varying its node size. Toponym importance is computed as the sum of all weighted relationships for a toponym with all other toponyms in the network. The larger the node, the more important is the toponym in the network. Similarly, the thicker a link between toponyms on the network, the stronger their inter-relationships.

Figure 2: Toponyms relationships in text space (a) and in geographic space (b).

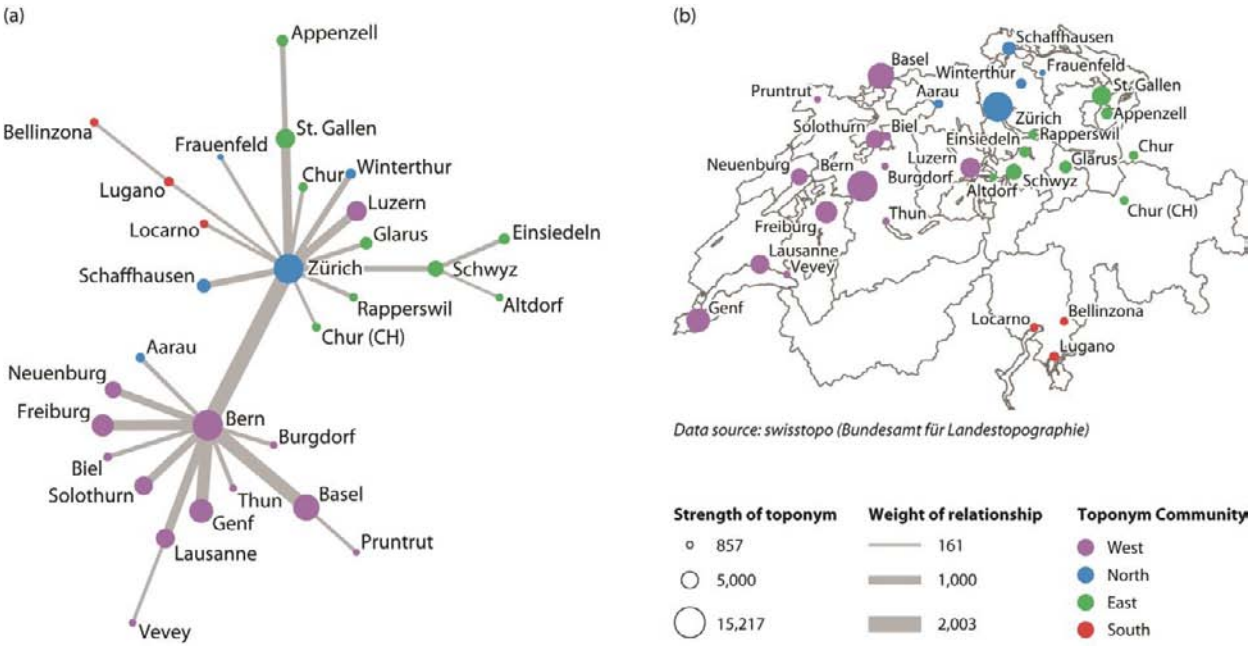


Figure 2b shows the spatial distribution of the extracted toponyms on a map of Switzerland. Again, node size represents the importance of the toponym, as explained above. We can now compare and contrast the extracted toponym patterns in the network spatialization and in the map. This allows for first visual inspection of the employed methods, and further validation and evaluation. A first striking result is that the geographic pattern (2b) is replicated in the network spatialization (2a). In other words, toponyms that are close in distance in geographic space are also closely related in text co-occurrence space. Two large nodes dominate the network in Figure 2a. With *Zürich*, currently Switzerland's financial center, and *Bern*, the country's political capital, two well-known cities in Switzerland are prominently depicted in the center of the network. They are both also strongly related. These two cities form a major axis in the network linking a *West*, *North*, *East*, and *South* toponym cluster. The bottom part of the network in Figure 2a shows *Bern* as a major hub for the *West* cluster, except for *Aarau*. Considering the upper part of the network in Figure 2a, the nodes connected to *Zürich*, except for *Luzern*, are all part of the *East*, *North* and *South* clusters. *Aarau* and *Luzern* look like outlier in the network space. The map in 2b might explain this: *Aarau* is almost equally far apart from the *West* community cluster and the *North* communities. While *Luzern* is geographically located in close proximity to many nodes in the *West* cluster, it has its strongest relationships with *Zürich*.

A center-periphery pattern is visible both in the network and in the map. The smaller towns *Bellinzona* and *Einsiedeln* are not directly connected to the large center nodes *Bern* and *Zürich*. They are connected to regional center nodes like *Lugano* and *Schwyz*. This might be an indicator of local spatial clustering, and it might suggest that the network spatialization not only reveals horizontal spatial relationships, but also reproduces a spatial settlement hierarchy.

The visualization also reveals that the toponyms *Biel* and *Altdorf* are wrongly located in the map. The city of *Chur* not only exists in Switzerland, but also in the *Principality of Liechtenstein* by mistake, as the Swissnames gazetteer also contains toponyms from the *Principality of Liechtenstein*. In other words, the visual analytics approach also supports us in identifying algorithmic problems. Improvements of the algorithm will be evaluated in future work.

5 Discussion

Interestingly, geographic distance has a strong effect on the relationships between toponyms extracted from a historic text database. Indeed, extracted node clusters exhibit a highly spatially auto-correlated pattern in the map, and in the spatialized network. Already in 1971 Tobler & Wineburg [27] predicted unknown locations of historic *Cappadocian* towns using a gravity model, based on co-occurrences of place names in old Assyrian records. They speculated that interactions between cities is proportional to their populations, that is, the larger the population of two cities, the more interactions occur between them. Similarly, following "Tobler's Law" [26] they postulated that if cities are mentioned together more often on a *Cappadocian* tablet, they must be closer to one another in geographic space, compared to cities that are mentioned less often. For the data set we analyzed in this study, [27]'s speculation seems to predict very well. Geographic distances seem to be of specific

importance in a historical dataset, as spatial separation was more difficult to overcome in the past than today. The concept of time-space convergence [13] which relates to the amount of space that can be covered in a given time period seems important to mention in this context. With the advancement of transportation technology people are able to cover larger distances in shorter amounts of time. In fact, the spatialized toponym network pattern reproduces the current transportation corridors in Switzerland. *Zürich* and *Bern* were then, and are still now, two major transportation hubs in Switzerland, perhaps because they are centrally located, and still today they are central nodes on a transportation network that links the Western and the Eastern parts of Switzerland. Other cities of high importance for Switzerland in the present as well as in the past such as *Basel* and *Genf* are not represented as central nodes in the network spatialization. One possible reason for this could be that relationships between these cities at the periphery might have especially in the past and also today been stronger with places located in neighboring countries than with cities in Switzerland, for instance, with places in France for *Genf* and *Basel*, and places in Germany for *Basel*. In our study, toponyms are only considered if they are located in Switzerland or in the Principality of Liechtenstein. An approach how to handle such edge effects can be found for example in [5].

6 Summary and Future Work

The aim of this study was to develop a framework based on GIR and GeoVA approaches to automatically uncover and visualize spatial, temporal, and thematic information buried in a digital dictionary about Swiss history (HDS). We showcase our approach focusing on the spatial information available in the text archive, and present a network spatialization based on co-occurrences of toponyms found in HDS articles. The uncovered network of toponyms illustrates the strong effect that geographic distance has on the historical relationships of places in Switzerland. The visual displays also helped us to uncover potential limitations of the employed GIR approach which we shall address in the future.

We are currently working towards an automated temporal analysis of the HDS articles to allow for change detection in the spatial structure and organization of toponyms in Switzerland's history. We also aim at assessing and visualizing thematic relationships (e.g., economy, politics, etc.) between toponyms in the HDS corpus, and how these might have changed over time. This may help to better explain the uncovered structure and strength of toponym relationships.

A further next step will be concerned with developing a dynamic and interactive user interface. We are currently evaluating various frameworks for online visualization including the Data Driven Documents (D3) technology [4], to complement the existing online HDS. D3, a JavaScript library, provides methods to create powerful and interactive visualization components for the Web. One component of the interface will allow users to query specific articles by space (i.e., through a cartographic map), by time (i.e., by selecting time slices), and by theme (i.e., using thematic filtering options). A second component of this interface will allow users to visually explore spatial, temporal, and thematic relationships by means of network spatializations and SOMs, as shown in Figures 1 and 2. Network spatializations might allow users to uncover hidden relationships between spatial entities over time, and, for example, how inter-city relationships might have changed over time. We envision dynamic network visualizations to emphasize *change* in the explored historical database. Thematic information will be re-organized using SOMs, which may serve as base layer onto which the change of events over time might be depicted. Spatial entities may be projected onto the self-organizing map as well, of course, and dynamically visualized, using the temporal information extracted from the text documents.

Acknowledgements

We would like to thank Curdin Derungs, Jannik Strötgen and Julian Zell who specifically helped us to implement the GIR part of our research. We are also grateful to Ross S. Purves and Damien Palacio for their invaluable feedback on this research project.

References

[1] O. Alonso, M. Gertz and R. Baeza-Yates. On the Value of Temporal Information in Information Retrieval. *SIGIR Forum*, 41(2): 35-41, 2007.

[2] O. Alonso, J. Strötgen, R. Baeza-Yates and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011)*, 2011.

[3] V. D. Blondel, J.-L. Guillaume and R. Lambiotte. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[4] M. Bostock, V. Ogievetsky and J. Heer. D3: Data-Driven Documents. In *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. http://d3js.org/ (March 2014).

[5] A. Bruggmann. Netzwerkvisualisierung der Ostschweiz. Zurich: University of Zurich, 2012.

[6] R. Burns and A. Skupin. Towards Qualitative Geovisual Analytics: A Case Study Involving Places, People, and Mediated Experience. *Cartographica*, 48(3): 157-176, 2013.

[7] D. Buscaldi. Approaches to Disambiguating Toponyms. *The SIGSPATIAL Special - Letters on Geographic Information Retrieval*, 3(2): 16-19, 2011.

[8] C. Derungs and R .S. Purves. From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, 2013. DOI: 10.1080/13658816.2013.772184.

[9] S. I. Fabrikant and A. Skupin. Cognitively Plausible Information Visualization. In J. Dykes, A. M. MacEachren and M.-J. Kraak, editors, *Exploring Geovisualization*, 667-690, 2005.

[10] B. Hecht and M. Raubal. GeoSR: Geographically Explore Semantic Relations in World Knowledge. In L. Bernard, A. Friis-Christensen and H. Pundt, editors, *11th AGILE International Conference on Geographic Information Science*, 2008.

[11] S. R. Hespanha and J. P. Hespanha. Text Visualization Toolbox - a MATLAB toolbox to visualize large corpus of documents. 2011. http://www.ece.ucsb.edu/~hespanha (March 2014).

[12] Historical Dictionary of Switzerland (HDS). 2014. http://www.hls-dhs-dss.ch/ (February 2014).

[13] D. G. Janelle. Spatial Reorganization: A Model and Concept. *Annals of the Association of American Geographers*, 59 (2): 348-364, 1969.

[14] T. Kohonen. Self-organizing maps. Springer, Berlin, 2001.

[15] J. L. Leidner. Toponym Resolution in Text. Doctoral Dissertation. Edinburgh: University of Edinburgh, 2007. http://hdl.handle.net/1842/1849 (February 2014).

[16] NWB Team. Network Workbench Tool 1.0.0. 2006. http://nwb.slis.indiana.edu (March 2014).

[17] S. Overell. The Problem of Place Name Ambiguity. *The SIGSPATIAL Special - Letters on Geographic Information Retrieval*, 3(2): 12-15, 2011.

[18] J. Pustejovsky, J. Castaño, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz and D. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, 28-34, 2003.

[19] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro and M. Lazo. The TIMEBANK corpus. In *Proceedings of corpus linguistics*, 647-656, 2003.

[20] M. M. Salvini. Spatialization von nutzergenerierten Inhalten für die explorative Analyse des globalen Städtenetzes. Doctoral Dissertation. Zurich: University of Zurich, 2012.

[21] A. Skupin and P. Agarwal. Introduction: What is a Self-Organizing Map? In P. Agarwal and A. Skupin, editors, *Self-Organising Maps: Applications in Geographic Information Science*, 1-20, 2008.

[22] A. Skupin and S. I. Fabrikant. Spatialization. In J. P. Wilson and A. S. Fotheringham, editors, *The Handbook of Geographic Information Science*, 61-79, 2007.

[23] M. Steyvers and T. Griffiths. Probabilistic Topic Models. In T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, 2007.

[24] J. Strötgen and M. Gertz. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2): 269-298, 2013.

[25] swisstopo. SwissNames. 2014. http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html (February 2014).

[26] W. Tobler. A Computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2): 234-240, 1970.

[27] W. Tobler and S. Wineburg. A Cappadocian Speculation. *Nature*, 231: 39-41, 1971.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Geo-Information Visualizations of Linked Data

Rob Lemmens
University of Twente
Faculty of Geo-Information Science and Earth Observation (ITC)
P.O. Box 217
7500 AE Enschede
The Netherlands
r.l.g.lemmens@utwente.nl

Carsten Keßler
Center for Advanced Research of Spatial Information and Department of Geography
Hunter College, CUNY
695 Park Avenue
New York, NY-10065
USA
carsten.kessler@hunter.cuny.edu

Abstract

Linked Data provides an ever-growing source of geographically referenced data for application development. In this paper, we analyse the workflow behind the development of such an application. Using two examples based on worldwide development aid and refugee data, we discuss the steps from locating data for use and data integration, up to the actual visualization in a web-based application. At each step, we discuss the skill set required for completion and point to potential challenges. We conclude the paper by putting our case study in the context of GIScience curriculum development.

Keywords: Linked Data, visualization, development, frameworks, workflow

1 Introduction

The amount of geographic information available as Linked Open Data (LOD) is rapidly increasing and becoming an invaluable source for application development. The term Linked Data refers to a set of best practices to publish machine-readable and semantically annotated data online [1]. The approach builds on established Web standards for identifying and accessing data sources (URLs), lightweight semantics (RDF) for data description, and a standardized query language for data access (SPARQL). These principles facilitate a distributed and interlinked collection of datasets known as the Linked Data Cloud [3]. Geographic information sources such as GeoNames[1] play a central role in this cloud, which is also documented by new datasets from cultural heritage [5], environmental monitoring [6], and emergency response [7], as well as the OGC GeoSPARQL query language [8].

At the same time, software development has transformed towards cloud environments and multi-platform development, especially including mobile devices. New software development platforms and libraries have eased the development of interactive web pages and mobile apps. Examples are web frameworks such as Django,[2] online content management systems such as Drupal[3] and mobile app platforms such as PhoneGap[4] and App Inventor[5] [10].

The goal of this paper is to review the process to get from LOD to a working application and put it in the context of the required skillset. We sketch the steps in developing web-based visualizations of humanitarian data (see Figures 1 and 2) and draw conclusions concerning the practical and conceptual skills that need to be covered in a GIScience curriculum for students to be able to complete such a development task.

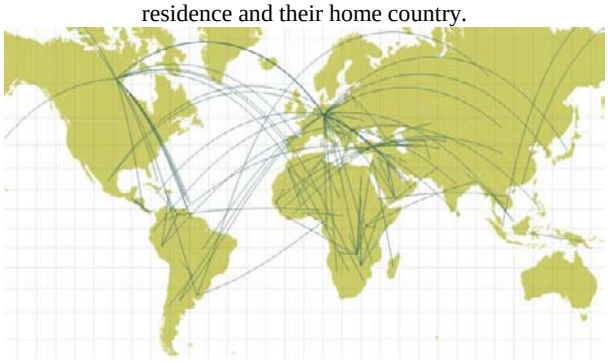
Figure 1: Web-based visualization of data from the International Aid Transparency Initiative. Brighter colors indicate higher amounts of development aid received.

[1] http://geonames.org
[2] https://www.djangoproject.com
[3] http://drupal.org
[4] http://phonegap.com
[5] http://appinventor.mit.edu

2 Development workflow

This section describes the different steps that were required to build the two sample web applications and discusses the different skills required to complete them. Figure 3 gives an overview of the different components and their interplay.

Figure 2: Web-based visualization of UNHCR refugee data. The blue arrows connect the refugees' current country of residence and their home country.



2.1 Locating data

Linked Data sets can be provided as RDF files in different formats or through SPARQL endpoints. Registries such as W3C SparqlEndpoints[6] and datahub[7] act as a good starting point to look for data relevant to the given application, from which the developer can look for related (i.e., *linked*) datasets. This process requires a general understanding of the Linked Data principles and potentially some proficiency in the SPARQL query language. In case of the two sample applications developed for this case study, the datasets included data from the International Aid Transparency Initiative,[8] the Humanitarian eXchange Language [7],[9] UNCHR refugee statistics (self-hosted), DBpedia,[10] and currency conversion rates.[11]

2.2 Data access

Whether the data is available as a file or from a SPARQL endpoint, data access will typically start by exploring the dataset, e.g., by listing the resources provided, or by browsing their types and the properties that describe them. This process iteratively leads to a query that generates the subset of the dataset the developer wants to process in her application, and it often includes reverting to locating additional data sources if information is missing.

Retrieving the data in the required form can also prove challenging. In both of our examples, the goal for the visualization was to show aggregates, i.e., the total amount of development aid that went to a given country, and the total number of refugees from country A that are currently in country B. The actual data, however, were highly disaggregated, e.g., by donor (IATI) or by demographic breakdown (UNCHR). The extra steps in the query require in-depth knowledge of the SPARQL query language and pose an additional challenge for novice developers. The following query, for example, asks for the total number of refugees from country A in country B, as specified in the UNHCR data:

```
prefix hxl: <http://hxl.humanitarianresponse.info/ns/#>
prefix dbpprop:    <http://dbpedia.org/property/>

SELECT DISTINCT ?fromCode ?toCode (SUM(?count) AS
?refugees) WHERE {

    ?pop  hxl:atLocation ?to ;
          hxl:placeOfOrigin ?from ;
          hxl:personCount ?count .

    ?to   hxl:atLocation ?country .

    ?country dbpprop:isoCode ?toCode .
    ?from    dbpprop:isoCode ?fromCode .

    FILTER (?count > 0)

  } GROUP BY ?fromCode ?toCode ORDER BY ?fromCode
```

While we have only worked with separate datasets for the visualizations presented here (option 1 in Figure 3), a fully *distributed* solution based on federated queries (option 3 in Figure 3) would require additional data. For the IATI application, for example, the development aid numbers provided as LOD are in different currencies, so they all have to be converted to a common currency. This requires an additional data source with currency conversion rates, such as currency2currency [12].

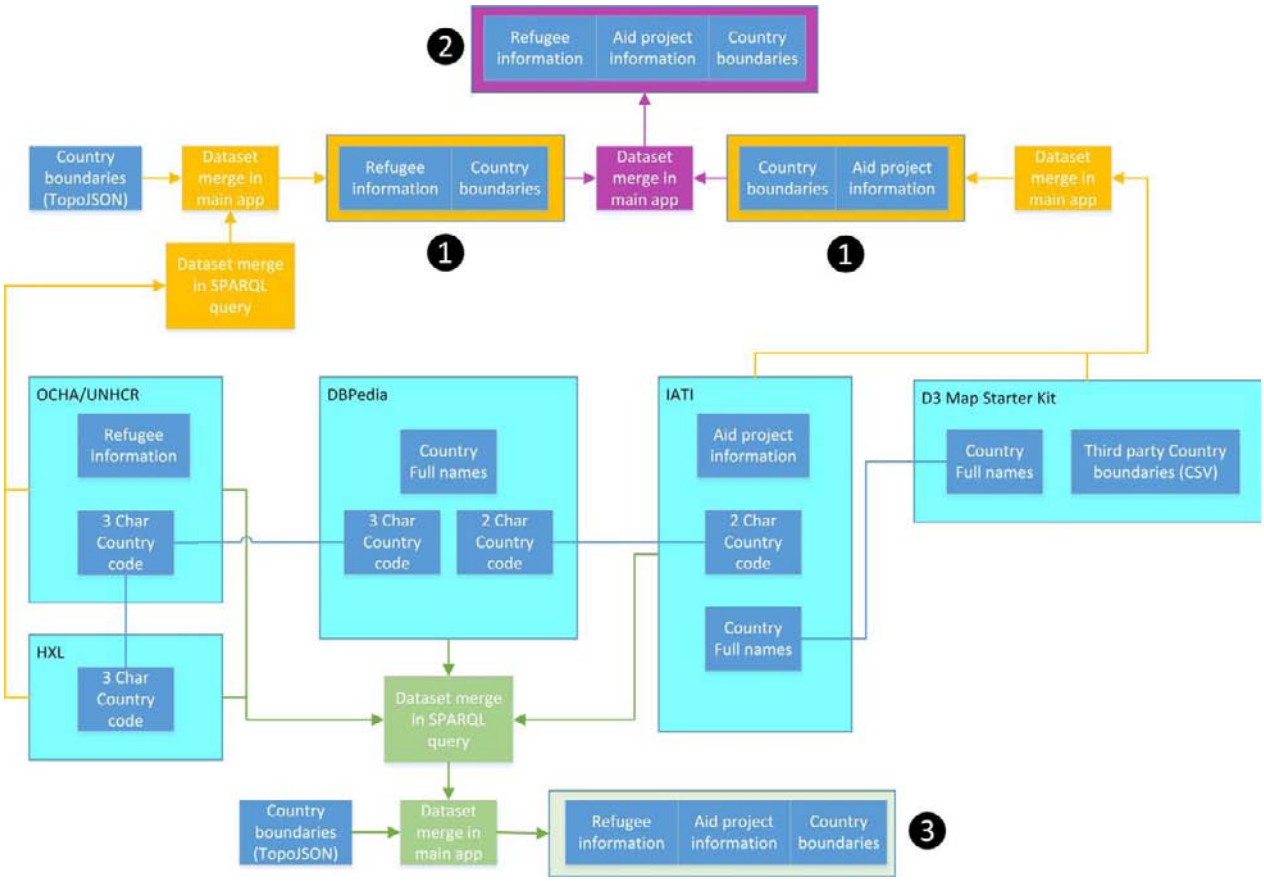
2.3 Data integration

Whenever more than one dataset is required for the application, these datasets in most cases have to be integrated in some way. If the goal is a simple visualization on a map, and the involved datasets include spatial references, the integration can be done on the map. In that case, this is a purely *visual* integration, and no further work is required.

In most cases, however, the underlying data will have to be integrated through common identifiers – similar to joins between tables in a relational database. In our IATI application, for example, we had to join the IATI data and the currency conversion rates to DBpedia, since the former uses 3-letter ISO currency codes, while the latter uses DBpedia URIs as identifiers for the currency codes. The corresponding integration can be implemented either in the query or in the application. An implementation in a *federated query* [9] that accesses multiple RDF datasets at once has the advantage that the result is a single file that can be directly processed by the framework used for the user interface. However, this approach is often slow since SPARQL results from multiple endpoints have to be collected, integrated, and returned to the client. Querying each dataset separately from the application is often faster, but results in multiple files that have to be integrated at the application level, thus placing more load on the client. Again, these considerations require knowledge about different querying and caching techniques to improve response time, depending on how frequently the queried datasets are updated.

[6] http://www.w3.org/wiki/SparqlEndpoints
[7] http://datahub.io/organization/lodcloud
[8] http://aidtransparency.net; data provided as LOD by VU Amsterdam at http://eculture.cs.vu.nl:1987/iati/home
[9] http://hxl.humanitarianresponse.info
[10] http://dbpedia.org
[11] http://currency2currency.org

Figure 3: workflow components and integration options.



2.4 Data output and visualization

While the XML-based SPARQL results format that endpoints return by default is very uncommon in any non-semantic web environment, the results can also be obtained in more common formats, such as CSV or JSON. The desired response format for a query can be set through an additional parameter in the HTTP request, or by setting the corresponding HTTP accept header. Both approaches require basic knowledge of the HTTP protocol and experience in using libraries such as cURL.[12]

The decision which results format should be chosen hinges on the input formats supported by the library chosen for the user interface. In a web development context, SPARQL query results can be shown by dedicated tools such as sgvizler [11] and Spark.[13] Web-based data aggregation tools such as Highcharts[14] and Google fusion tables[15] allow for combining spreadsheet-type information into graphs and simple maps on the web. Geo-information representation tools such as OpenLayers[16] and Leaflet[17] specifically handle georeference systems and map rendering.

We have opted for a generic and scalable tool based on Javascript: the D3 (Data Driven Documents) library, as this provides powerful capabilities for all of the above, is fairly easy to learn and is well documented. Any of the options listed above requires a certain level of proficiency in JavaScript, HTML, and CSS. While we focus on web-based applications here, developing native applications for desktop or mobile platforms adds another level of complexity.

3 Application: IATI data visualization

To demonstrate the needs for the abovementioned app development, we take the use case of creating web-based visualizations of humanitarian data, coming from different sources. In our case, these sources can be combined in different ways, basically through SPARQL queries and by data merging in the app. The latter is implemented by D3 JavaScript functions.

Figure 3 depicts the options: Separate visualizations (1), combined visualization by app merge (2) and combined visualization by SPARQL query (3). Combined visualizations allow for an integrated analysis of sources. In contrast to studies such as Findley et al. [4], which demonstrates geographical correlations between foreign aid and armed conflicts, we do not intent to explain such correlations, but rather focus on the technical aspects of data source integration.

[12] http://curl.haxx.se
[13] http://www.revelytix.com/content/spark
[14] http://www.highcharts.com/
[15] http://www.google.com/drive/apps.html#fusiontables
[16] http://openlayers.org/
[17] http://leafletjs.com/

The International Aid Transparency Initiative (IATI) fosters the exchange of information on international aid projects. IATI does this by setting standards for information exchange and providing a hub for registering data sets. IATI does not provide the data itself, this is done by the donor organizations themselves. IATI does provide information about how to create and consume IATI-standardized information and about available tools by third parties. IATI information has been deployed in a triple store [2] and is available as a SPARQL endpoint.

Since a federated query approach proved too slow during the data integration step, subsets of the used datasets were exported using SPARQL construct queries and loaded into a local triple store. This allowed for faster iterations during the development of the integration query, which was then ultimately used to produce a CSV file fed into D3. In a production environment, this file could be produced directly from the original endpoint via federated query and cached, with updates e.g. on a daily basis, depending on the data update frequency.

4 Conclusions

The amount of Linked Open Data containing geographic information is growing and becomes an attractive data source for application development. Based on the premise of truly *linked* data, it should be straightforward to use data from different sources together in applications. In reality, the integration of data from such sources to be able to use them together is still challenging, leading to situations where it is easier and more straightforward to download subsets of the data and integrate them locally. While this is a practice-oriented approach, it is clearly not in the spirit of Linked Open Data.

Once the data for an application has been assembled, the developer is confronted with the choice from a wide variety of frameworks for implementation. While many frameworks such as D3 have sophisticated functionalities for the visualization of and interaction with geoinformation, putting them to use still confronts novice developers with a steep learning curve. In order to implement the (relatively simple) visualizations shown in this paper, profound knowledge of RDF, SPARQL, HTTP requests, HTML, and JavaScript is required. Adding interaction and developing for touch screen devices, for example, adds another layer of complexity.

While adding all of these technologies to the already demanding GIScience curricula is hardly possible, we believe that the study programs can enable their students to learn these (and other) new technologies faster. Proficiency in different geo-information standards is already part of the curriculum in many programs and can easily be extended to a broader range of web standards. Existing research methods courses can be extended with sections on research for software development to familiarize students with resources such as StackOverflow[18] as well as tools such as GitHub[19] and bl.ocks.org.[20] Finally, hands-on lab exercises that ask for the development of creative solutions, rather than following "click-through" instructions, get the students used to independent problem solving.

[18] http://stackoverflow.com
[19] https://github.com
[20] http://bl.ocks.org

References

[1] Berners-Lee, T.: Linked Data – Design Issues (2009) Online: http://www.w3.org/DesignIssues/LinkedData.html

[2] Brandt, K. (2013), Linked Data for IATI, MSc Thesis,Vrije Universiteit Amsterdam.

[3] Cyganiak, R., Jentzsch, A. (2011) Linking Open Data cloud diagram. Online: http://lod-cloud.net

[4] Findley, M. G., J. Powell, D. Strandow, and J. Tanner (2011), The Localized Geography of Foreign Aid: A New Dataset and Application to Violent Armed Conflict, *World Development*, *39*(11), 1995–2009, doi:10.1016/j.worlddev.2011.07.022.

[5] Haslhofer, B., & Isaac, A. (2011). data.europeana.eu: The Europeana Linked Open Data Pilot. In International Conference on Dublin Core and Metadata Applications (pp. 94-104).

[6] Kauppinen, T., de Espindola, G. M., Jones, J., Sánchez, A., Gräler, B., & Bartoschek, T. (2013). Linked brazilian amazon rainforest data. Semantic Web.

[7] Keßler, C. and Hendrix, C. (forthcoming) The Humanitarian eXchange Language: Coordinating Disaster Response with Semantic Web Technologies. Semantic Web Journal, accepted.

[8] OGC (2012) GeoSPARQL – A Geographic Query Language for RDF data.

[9] Prud'hommeaux, E.,Buil-Aranda, C. (2013) SPARQL 1.1 Federated Query. W3C Recommendation: http://www.w3.org/TR/sparql11-federated-query/

[10] Shih, F., O. Seneviratne, D. Miao, I. Liccardi, L. Kagal, E. Patton, C. Castillo, and P. Meier (2013), Democratizing Mobile App Development for Disaster Management, in *AIIP '13 Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities*, pp. 39–42, ACM.

[11] Skjaeveland, M. (2012), Sgvizler: A javascript wrapper for easy visualization of SPARQL result sets, in *Extended Semantic Web Conference*.

[12] Stolz, A. and Hepp, M.(2013) Currency Conversion the Linked Data Way, in: Proceedings of the Workshop on Services and Applications over Linked APIs and Data (SALAD2013), in conjunction with the 10th Extended Semantic Web Conference (ESWC 2013), May 26-30, Montpellier, France.

Session:
Geospatial Algorithms

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Estimating Moving Regions out of Point Data – from Excavation Sites in the Amazon region to Areas of Influence of Prehistoric Cultures

Carolin von Groote-Bidlingmaier
University of Augsburg/Institute for Geography
Alter Postweg 118
86159 Augsburg, Germany
cvgb@geo.uni-augsburg.de

Sabine Timpf
University of Augsburg/Institute for Geography
Alter Postweg 118
86159 Augsburg, Germany
Sabine.timpf@geo.uni-augsburg.de

Klaus Hilbert
University of Porto Alegre (PUCRS)/FFCH
Av. Ipiranga, 6681 – Bairro Partenon
Porto Alegre, Brazil
hilbert@pucrs.br

Abstract

How can we derive the changing area of influence of specific cultures from only a few excavation sites in the Amazon region? The approach used for calculating areas of influence for several time intervals strongly depends on the kind of available input data and the examined issues. Our approach divides the input point data into different time intervals and calculates an area (or areas) of influence for each, factoring in spatial and temporal uncertainties inherent in the data. The computation is based on a cost surface, which is derived from the needs and capabilities of the analyzed prehistoric culture or tradition. To take into ac-count that archaeological data is inherently vague, the database is able to handle spatial uncertainties by applying varying maximum distances. Based on the cost raster and the maximum distance a maximum cost value is calculated which is used to derive the said area(s) of influence, which can then be analyzed for changes.
Keywords: Moving Objects, Point to Area, Archaeology, Spatio-Temporal Uncertainty.

1 Introduction

This research is about the estimation of moving regions out of point data sets taking into account specific spatio-temporal relationships in the data. In concrete terms we mean to calculate area(s) of influence out of excavation data in the Amazon Region.

A lot of research focuses on moving points, which, in most cases may be satisfactory [11]. For other approaches the use of points for the representation of moving objects is not adequate and further research is necessary, e.g. when analyzing the changes of areas between two or more time steps, especially, when geographical aspects should be factored into the analysis.

This research focuses on the problem of deriving moving regions from temporal point data. This case study copes with very little point data sets and even missing values. The difficulties we had in establishing a time orientated movement patterns of the prehistoric indigenous cultures throughout the great Amazon area were, mainly related to the often missing radiocarbon dates.

This work focuses on the macro level to understand the spatio-temporal processes and relations of cultures in the whole Amazon Region.

In order to reduce the temporal uncertainty caused by the lack of radiocarbon data this research is based on the following hypothesis:

- The first appearance of a new culture starts with settlements at strategically important sites which are big and used by more than one culture (so called multi phased excavation sites).
- This is followed by a period of expansion and the foundation of new settlements. Next to the important multi phased sites there are several others (still big) which were only used by a specific culture (single phased).
- The era of the culture ends with retreat, shrinking and smaller settlements. Which means that only small single phased excavation sites are factored.

This distinction allows estimating three time intervals. The few given radiocarbon dates are used to specify the transition period.

The rest of the paper is organized as follows: in the second chapter a brief over-view over the existing literature is given. Chapter three describes the method which is used to calculate the area(s) of interest. The case study is shown in chapter four and the conclusion and discussion are presented in chapter five.

2 Background

There is no coherent definition of archaeological cultural divisions. We understand tradition as a group of elements or techniques with temporal persistence at a larger scale, and culture as any material culture complex and features related in time and space at a more regional or site-related scale [15].
The idea of deriving regions from archaeological site patterns is not new and has been commented on before.

"The ability to establish territorial extents of political, religious or economic zones in a quantitative, formally correct way allows us to transfer hypotheses and knowledge from observations made at single archaeological sites to the landscape surrounding them – effectively moving from point to area-based descriptions." ([4] p. 245). A "territory" is supposed to be explored by inhabitants of the settlements and its boundaries are defined by "low-cost" factors and by the maximum circumference people are willing to walk in order to get food or other resources [18]. In order describe the political influence (I), [16] presented the so-called Xtent model. It factors in the settlement sizes and distances between them. Two coefficients determine the weight of the parameters and therefor the balance between size and distance.

[4] presented an enhanced version of the Xtent model, which allows the integration of topographic features. Both versions are highly dependent on the used coefficients due to the underlying formula invented by [16].

Another approach to model areas of influence is the site catchment analysis. It is based on the supposition that every settlement is surrounded by a catchment area understood as a zone of resources, domestic or wild, inside an easy travelling distance from the settlement. The further the resources from the site the more difficult their exploitation. Therefore the analysis needs specific knowledge about the excavation site and draws conclusions based on the findings. The size of the areas is determined by the sources of the materials (e.g. stones which don't exist close to the Amazon River) or the maximum traveling distances to procure resources [18].

3 How to calculate an area of influence

In our approach, a combination of the territory and site catchment analysis is being used to calculate the area of influence. The maximum territory is estimated using a maximum distance a culture is willing to walk. On the basis of this, the site catchment is calculated which factors in the distances to resources needed by the culture (e.g. stones).

3.1 Database

The core part (orange) holds the information about the excavation sites and their findings. Each excavation site is unique but it can contain several findings, which in turn can be assigned to different cultures and traditions. The thematic part on the lower right in figure 1 (red) stores the cultures and traditions and the relation between them. The part handling the information about the maximum spatial extent (and by that the importance of a certain excavation site for one tradition) and its spatial uncertainty is displayed on the upper right (green). Based on the fundamentals of fuzzy-logic a truth value alpha (α) is used to factor the spatial uncertainty. That means the higher the α-value, the more exact is the maximum distance for that specific excavation site and culture. The fourth part (brown) handles the temporal information and is basically one table which stores the radiocarbon data and its uncertainty interval (e.g. 450±45 BP). The last three tables (blue) are used to store metadata. Besides the already mentioned data, some meta-information about the archaeologist, the source of the data and the institute, which performed the radiocarbon dating, is being collected (see figure 1 for complete database schema).

3.2 Creation of C_{rast}

The point data from the excavation sites form the basis for the analysis. Additionally, environmental data about rivers and the location of waterfalls is integrated.

The needs and capabilities for each culture depend on environmental, social and economic circumstances. The soil quality or trading relations are only two variables, which may have an influence on the shape of the area.

The carrying capacity of an area is understood as the number and density of people that any region can support [5]. The carrying capacity of an area can be altered by people manipulating their land by burning wood or producing fertile soil, planting trees and crops [2]. Therefore, a cost raster (C_{rast}) is calculated, which provides the basis for the site catchment creation. The natural resources in the Amazon (besides water and its related occurrence of food) are quite similar, thus they are considered to be ubiquitous.

As input for C_{rast} several information sources may serve, which are relevant for settlement in the Amazon Region. Based on [3] settlements are usually close to rivers due to the need of water. Therefore, an inverse cost distance raster is calculated for the rivers. The blackwater rivers (less nutrients, fish, etc.) gradient increases faster than the nutritious whitewater rivers's gradient. The flow direction of the Amazon River (not its tributaries) is included. In addition, the distance to the waterfalls is factored in, because this is the nearest location to get stones from.

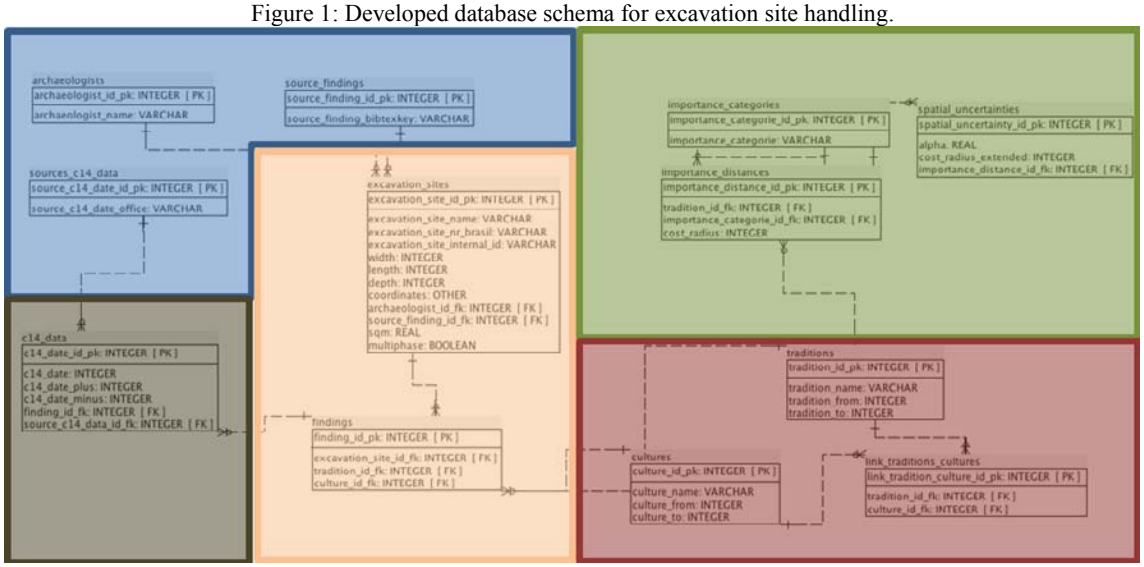
3.3 Deriving the area(s) of influence

C_{rast} serves as input for the calculation, which returns the cumulative costs between the input locations (the excavation sites) and every other cell. In the literature the maximum distance a culture can walk to procure resources varies a lot. In the Amazon Region some required goods (e.g. stones) are not available close to the Amazon but at the waterfalls in the inland. This extends the area of influence and it has been hypothesized that up to 25 kilometers are realistic for cultures in the Amazon Region [6, 10].

A maximum distance value ($Dist_{max}$) – which is provided by the database, and can be extended using a smaller α-value – is used to calculate the maximum cost distance value (CD_{max}) for each excavation site. That is done by:

1. buffering each excavation site with its maximum distance value $Dist_{max}$.
2. calculating the least cost path between the excavation site and the outline of the buffer (see figure 2). The resulting accumulated least cost path value equals the value CD_{max}

Figure 1: Developed database schema for excavation site handling.



3. creating the area(s) of influence on the basis of C_{rast}. Therefore the cumulative cost value CD_{max} is used as maximum cost value.
4. merging the resulting rasters to represent the connected area(s) of influence for one culture during one time interval (see figure 4).

Based on the given data, a separation in three time intervals for each culture is being made (see above) which are defined as follows:

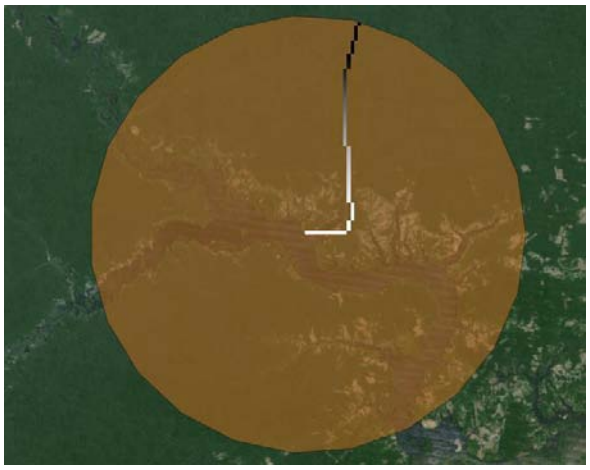
- ti_1: mp (multiphased)
- ti_2: sqm>=min(sqm(ti_1))
- ti_3: ¬mp&sqm<min(sqm(ti_1))

with ti_1 indicating the beginning of the settlement, ti_2 representing the interval with the widest spread (which also means a spatial overlap between ti_1 and ti_2) and ti_3 showing the time of shrinking. With regard to Allen's temporal relations [1], the time intervals defined above are not necessarily in a ti^-_1 "meets" ti^+_2 relation. The only mandatory condition is that ti_1 starts before or at the same time than ti_2 and that it ends before or at the same time than ti_2 (the same conditions are valid for the relation between ti_2 and ti_3). Besides "meets" also "equals", "starts", "finished-by" and "overlap" are possible relations.

The chronological order of the time intervals is known, whereas the point of transition between the different time intervals is not known. The transition period (Sep_{tix_tix+1}) between the time intervals can formally be described using radiocarbon data (see figure 3 with RC = radiocarbon, mp = multi phased, sp = single phased, LKD and HKD = least/highest known date, and sqm = squaremeter).

To factor spatial uncertainty a range of $Dist_{max}$ values is defined. This is handled by the database in which predefined α-values and associated extended $Dist_{max}$ values are stored. The extended $Dist_{max}$ values can be set independently for each finding, which allows individual treatment of each finding and therefore a higher accuracy for the findings which are well known. This is similar to the fuzzy spatiotemporal information system for handling excavation data introduced by [19].

Figure 2: Calculating the least cost path – and therefore CD_{max} – for an excavation site and a predefined $Dist_{max}$



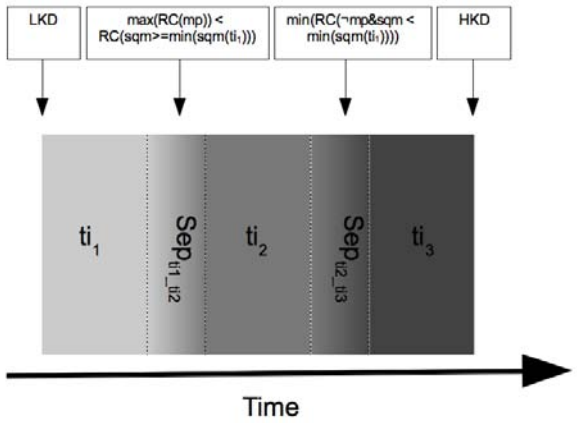
4 Case Study

The case study focuses on a culture named Guarita which is classified as "Polychrome Painting" tradition and is known to have settled in the Amazon Region approximately between 600 AD and 1300 AD [8].

The necessary data (excavation site size, coordinates, and number of traditions) are available for almost all excavation sites in our database, which makes them most suitable. Unfortunately, only very little radiocarbon data exists, therefore no temporal progress can be calculated.

The area(s) of influence were calculated using three different α-values – 1, 0.8 and 0.5 – for each excavation site. The results can be seen in figure 4.

Figure 3: Time line which shows the chronological order and formal description of the settlements.



The culture is located in the confluence area between the Rio Negro, Rio Solimões, the Rio Madeira and the Amazon River, the distribution pattern of the pioneer settlements of the Guarita Culture show four separated clusters. The main habitation sites are located on top of the bluff zones of the "terra firme", close to the fertile alluvial plains (várzea) and in contact with both river types.

The second stage shows an almost explosive increase of new habitation sites, away from the primary areas and following up the Rio Negro (north west), as well as the Madeira River (south east). Although the first is a blackwater river and the second a whitewater river, there is apparently no variation in the strategy of occupation of these ecologically different regions.

The third time interval is characterized by abandonment and retreat. All archaeological sites along the Madeira River were uninhabited, just like those located on the Lago Silves. Many sites at the mouth of the Rio Negro were abandoned as well.

5 Conclusion

The goal of this research is to determine the changes or movement of influence areas of precolonial cultures in the Amazon region using archaeological excavation data and assumption on the environmental conditions. One big problem to cope with when estimating an area of influence for pre-colonial cultures is the lack of data. There are only very few (excavation) points and it can be assumed that the used set of points is incomplete. The shown approach creates estimated area(s) of influence derived from point data. Due to the inherently vague spatial and temporal data, the areas are derived on a basis of maximum costs a culture is willing to overcome for procuring resources. The results seem to confirm the assumptions that after a period of expansion there is a shrinking process and a retreat to the backlands. The database schema as well as the area of influence calculation is not necessarily connected to the data of the Amazon Region but can also be applied to other archaeological datasets.

Spatial vagueness is factored in using different α-values, which have an influence on the selected buffer sizes for each excavation site. The temporal information is derived using specific properties of the sites. The results vary due to the chosen α-values: the higher the value, the more unsteady (but wider) are the calculated areas. By assigning the values for each excavation site the uncertainty can be reduced, because only the areas of the less known places extend.

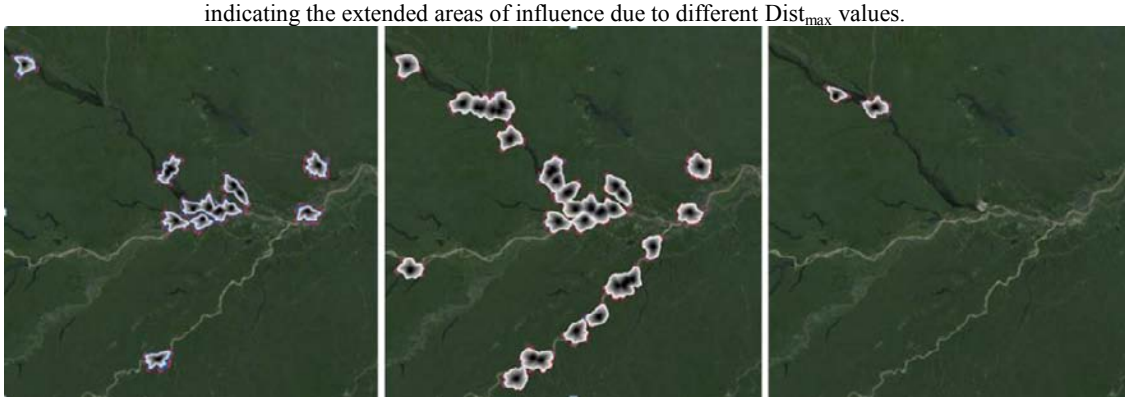
In the current stage of research, radiocarbon data is included for defining the separation interval between two time intervals. This is based on the problem that very few radiocarbon data is available for the examined cultures. Even though radiocarbon data is not exact (e.g. bias due to volcanic activity) it helps to specify the time and duration of the settlements.

The database is in the process of being completed with (analog) data published in archaeological and ethnological research papers. This collection process will also increase the amount of available radiocarbon data, allowing for an analysis of temporal occurrences. Fuzzy zones of change can be derived, where the phase changes between two time intervals. Thus, a better knowledge about the chronological sequence and settlement behavior can be achieved.

It must be assumed that the results will be distorted due to missing excavation sites. Some excavation sites with Konduri findings are not yet in the database due to missing additional information. They are located downstream of the Amazon River and were documented by Nimuendajú in 1937 [14].

The quality of the resulting areas is highly dependent on the used cost surface. As [9] claimed, the data must be collected carefully to assure a more realistic model of the factors, which determine the area. That means, the final output is only as

Figure 4: Results for the Guarita culture. Time intervals 1, 2 and 3 (from left to right). The red and blue colors are indicating the extended areas of influence due to different $Dist_{max}$ values.



good as the cost raster, which is also a matter of resolution. In addition, if some important resource or other input factor is unconsidered, the calculated area might not be convincing.

The migration behavior of pre-colonial cultures in the Amazon Region is widely debated. The main question is if the Amazon Region is an environment of abundance and therefore appropriate for permanent settlement (model of abundance [7, 12, 17] or an environment of limited resources [13] Our current approach is based on the 'model of abundance' theory, while being aware of the fact that other movement behaviors might have been predominant.

In the future, the method will be refined to cope with missing values and to derive more information about the connectivity between the settlements of one culture. One step leads to hypothesizing further (as yet unexcavated) sites by determining their suitability and assuming a settlement at the most suitable places. Another step is to use the calculated area of influence itself as a cost raster to calculate the least cost paths between the settlements to derive at a settlement network. This in turn could lead to a better estimation of costs because a directed graph can be used (faster movement on rivers but dependent on the flow direction etc.). Such a model can be extended to include other cultures, which existed at the same time interval, in order to analyze trading relations.

References

[1] J. F. Allen. (1983) Maintaining Knowledge About Temporal Intervals. *Communications of the ACM* 26(11): 832-843, 1983.

[2] W. Balée. *Cultural Forests of the Amazon. A Historical Ecology of People and their Landscapes*. University of Alabama Press, 2013.

[3] W. M. Denevan A Bluff Model of Riverine Settlement in Prehistoric Amazonia. *Annals of the Association of American Geographers* 86:654-681, 1996.

[4] B. Ducke B and P. C. Kroefges. From Points to Areas: Constructing Territories from Archaeological Site Patterns Using an Enhancend Xtent Model. In: *Layers of Perception. Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA). Berlin, Germany, April 2–6*, Berlin, 2007.

[5] B. Fagan. *In the Beginning. An Introduction to archaeology*. Scott, Foresman and Company, Glenview, 1988.

[6] M. Heckenberger. *The Ecology of Power. Culture, Place, and Personhood in the Southern Amazon, A.D. 1000-2000.* Routledge, London, 2005.

[7] M. Heckenberger and G. E. Neves. Amazonian Archaeology. *Annual Review of Anthropology* 38:251-266, 2009.

[8] P. P. Hilbert. Archäologische Untersuchungen am mittleren Amazonas: *Beiträge zur Vorgeschichte des südamerikanischen Tieflandes. (Marburger Studien zur Völkerkunde)*. Reimer Verlag, Berlin, 1968.

[9] I. Hodder I and C. Orton. *Spatial Analysis in Archaeology*. Cambridge: Cambridge University Press, Cambridge, 1976.

[10] T. Koch-Grünberg. Zwei Jahre bei den Indianern Nordwest-Brasiliens. Strecker und Schröder, Stuttgart, 1921.

[11] P. Laube. Progress in Movement Pattern Analysis. In: *Behaviour Monitoring and Interpretation – BMI – Smart Environments*. CRC Press, London, p 368, 2009.

[12] D. Lathrap *The Upper Amazon. Ancient peoples and places*. Thames & Hudson Ltd, London, p 256, 1970.

[13] B. J. Meggers. *Amazonia. Man and Culture in a Counterfeit Paradise*. Smithsonian Institution Press, Washington, 1996.

[14] H. C. Palmatary. The archaeology of the lower Tapajós Valley, Brazil. *American Philosophical Society*, Philadelphia, 1960.

[15] Programa Nacional De Presquisas Arqueológicas (PRONAPA). Terminologia arqueológica brasileira para a cerâmica. *Manual de Arqueologia No. 1*. Curitiba, 1966.

[16] C. Renfrew and E. Level. Exploring Dominance: Predicting Polities from Centers. In: *Transformations: Mathematical Approaches to Culture Change*, Academic Press, New York, p 515, 1979.

[17] A. C. Roosevelt. *Moundbuilders of the Amazon. Geophysical Archaeology on Marajo Island, Brazil.* Academic Press, New York, p 480, 1991

[18] D. Roper. The Method and Theory of Site Catchment Analysis: A Review. In: *Advances in Archaeological Method and Theory*, Vol. 2 (1979), p 119-140, 1979

[19] A. Zoghlami. Through a Fuzzy Spatiotemporal Information System for Handling Excavation Data. In: *Bridging the Geographic Information Sciences. International AGILE'2012 Conference, Avignon (France), April, 24-27, 2012.* Lecture Notes in Geoinformation and Cartography, Springer, Heidelberg, p 450, 2012

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

An algorithm for segmenting a feature set into equitable regions

Md. Imran Hossain
University of the Bundeswehr Munich
Institute for Applied Computer Science
Werner-Heisenberg-Weg 39
85577 Neubiberg, Germany
Imran.Hossain@unibw.de

Wolfgang Reinhardt
University of the Bundeswehr Munich
Institute for Applied Computer Science
Werner-Heisenberg-Weg 39
85577 Neubiberg, Germany
Wolfgang.Reinhardt@unibw.de

Abstract

A set of geographic features of the same class representing a geographic area is often required to be divided in to several subsets/regions so that the sum of a numeric attribute of the features in each subset/region remains almost equal and the bounding polygon of regions do not overlap with each other. This kind of non-overlapping regions formation with similar collective feature value is of great importance especially in the field of optimization and spatial decision support. The paper presents a novel algorithm to solve the above mentioned spatial analysis work. The algorithm is further implemented, tested and the results are discussed.

Keywords: Spatial algorithm, Spatial Analysis, Vector Segmentation, Equitable Region

1 Introduction

Segmentation of a feature set of the same class into several equitable and non-overlapping[1] regions depending on a feature property is often required especially in the domain of optimization and spatial decision support. For example, an evacuation assistance providing authority having 4 emergency evacuation units (vehicles) might be interested to divide the whole emergency area into 4 parts in a way that each part consists of approximately the same number of evacuees and the bounding polygon of the parts do not overlap so that the evacuation process ensure optimization. Another practical need of such segmentation could be apprehended with the scenario that a service provider wants to cover an area (neighbourhood) with service centres of same capability. Let us assume that with the limited resources the service provider can provide only 4 service centres to cover the neighbourhood of 205 peoples. In such a case the service provider will be interested to segment the whole area into 4 regions so that each region consists of approximately the same number of people (in this case ~ 51) so that the service centres operates in an optimized way. Beyond the mentioned examples, a number of other different application areas could be found where the need of such a segmentation of a feature set is inevitable. A more precise definition of the problem that we address in this paper is given below.

A geographic area G defined by a feature set consisting of n number of features with a numeric attribute A has to be completely divided into N ($N \in \mathbb{N} \mid 2 \leq N \leq n$) number of subsets/regions based on following 3 criteria.

Criteria 1: Each region should consist of a certain number of complete features of the feature set. A splitting of a feature is not allowed.

Criteria 2: The sum of the value of A of all features in any region R_i must be equal to $T \pm d$ [where T is calculated by summing up the values of A of all n features of the geographic area G and then divided by N and d is a deviance]. Maximum value of d is equal to the maximum value of A of any given feature within G. The deviance d has to be considered as a splitting of the features is not allowed. Besides, as it may not possible in all cases of given data sets that the value of the sum of A of all regions is within $T \pm d$, the number of regions not following the criteria has to be minimized.

Criteria 3: The bounding polygon of any region should not overlap with any other region means it can only touch other or/and remain as disjoint.

The main goal of this paper is to present a novel algorithm (section 3 for more detail) to solve the problem stated herein. The authors developed the algorithm and successfully implemented it with c# and ArcObjects library. Implementation of the algorithm and the results of its application are discussed in section 4.

2 Related works

Automated zone design (AZD) or regionalisation is a technique for which Shortt [2] has given the overview of its concept, terminology and methods. AZD is an umbrella term for quite a number of approaches to create zones from a set of basic blocks following given criteria. Among the automated zone design algorithms automated zone design procedure (AZP) is the most popular and widely used one. It was introduced by Openshaw [6, 7]. The AZP has been enhanced by Openshaw and Rao [8], Alvanides [4] and Alvanides et al [3]. Cockings et. al. [5] used automated zone design techniques to dynamically maintain existing zoning systems. There are also a lot of other application of AZP algorithm such as climate zoning, location optimization and many more. The AZP algorithm iteratively combinines and recombines sets of blocks in order to create output zones which are optimised based on a set of pre-specified design criteria [8].

AZP is not applicable to our task described in the introduction as firstly, AZP is applicable only to continuous and connected feature sets whereas in our case continuous and

[1] Non-overlapping regions means the boundaries of the regions are disjoint or/and in touch with each other.

discrete feature sets must be treated. Secondly in AZP a zone can exist in a disconnected multi-polygon form which means a zone's bounding polygon may intersect with other zone's bounding polygon which is prohibited in our case. Also it is required in our approach that the bounding polygon of each region must not overlap with any other region.

The territory design tool of ESRI [1] offers functionality to create, automatically balance, and maintain territories. The tool establishes potential franchise areas and assigns sales territories consisting of multiple variables and levels. Again, the territory design tool works on continuous and connected feature sets. Manual intervention is often required to make all territory balanced. In contrast to the territory design tool, our goal is to balance the regions (territory in territory design tool) automatically and to cover discrete feature sets as mentioned earlier.

3 The algorithm

Firstly, the input, output and criteria of the algorithm are defined hereafter.

Input:
- Geographic area $G = \{f_n \mid f_n \in F$ (set of features), f_n has a numeric attribute $A\}$
- N ($N \in \mathbb{N} \mid 2 \leq N \leq n$) = number of required subsets of G, N has to be defined by the user.

Output:
- N number of subsets R_n (subsets/regions)

Criteria:
- Region cannot be formed with splitted feature means a feature of a region is not allowed to be in a form like $f_i/m \mid m \in \mathbb{N}$.
- The sum of $|A|$ ($|A|$ is the value of attribute A) of any region defined herein with $O(R_i) = T \pm d$
- The bounding polygon of any subset $BNDline(R_i)$ do not overlap with the bounding polygon of any other.

The value of T is calculated by equation 1and the value of d is an element of set D. The value of d can ranges from 0 to the maximum value of $|A|$ of a given feature set G (equation 2).

$$T = \frac{\sum_{f=1}^{fn} |A|(G)}{N} \ldots \qquad (1)$$

$$d \in \mathrm{D} = \{\mathrm{q} \in \mathbb{Q} \mid 0 \leq q < MAX\ (|A|(G))\ \} \ldots \qquad (2)$$

In general, the algorithm prioritizes forming regions along the bounding line *BNDline(G)* of the input feature set G. This approach prevents features being unclassified and also prevent big differences among the regions. A region R_i is formed by grouping features around the bounding line until $O(R_i) = T \pm d$. Once no region formation is possible along the *BNDline(G)*, another bounding line is created for the features which are not classified into regions and regions are again formed along the new bounding line. This process continues until *N-1* regions are formed. The N^{th} region is formed with remaining unclassified features after formation of $N\text{-}1^{th}$ region and consequently it´s possible that the sum of $|A|$ may not be within $T \pm d$ in this case. The algorithm is described in more detail through the following steps.

Step1. Objective function: Objective function returns the value of T on which each region is formed. The algorithm of the objective function is given by figure 1.

Figure1: Objective function algorithm

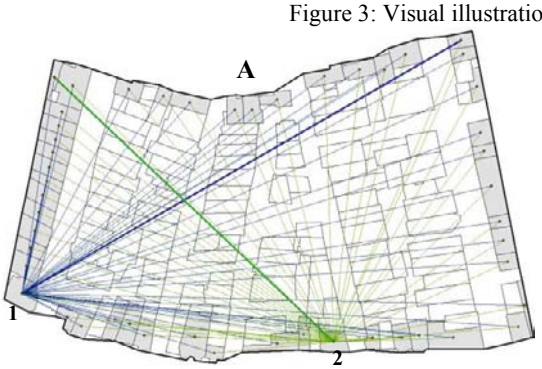
```
Data: Input FeatureSet G, Attribute A and No. of expected region N
Result: A double value representing T
1   FeatureSetTotal ← 0
2   for each Feature ∈ FeatureSet do
3       FeatureSetTotal = FeatureSetTotal + |A|
4   return FeatureSetTotal/N
```

Step2. Selection of the starting feature for the first region: The starting feature for the first region is selected by two criteria. Firstly, it has to be along the *BNDline(G)* and secondly it has to be located in an appropriate corner of G. Therefore the starting feature is selected by firstly making an array of features that touches the *BNDline(G)*. Secondly a feature is picked up from that array and distances are calculated from that feature to all other features of that array. The maximum distance is then stored with each picked up feature. This process is carried out for all features in the array. Finally, the feature that has the maximum distance value compared to other features in the array is selected as a starting feature for the first region formation (fig. 3A). The algorithm of this step is given in the following figure.

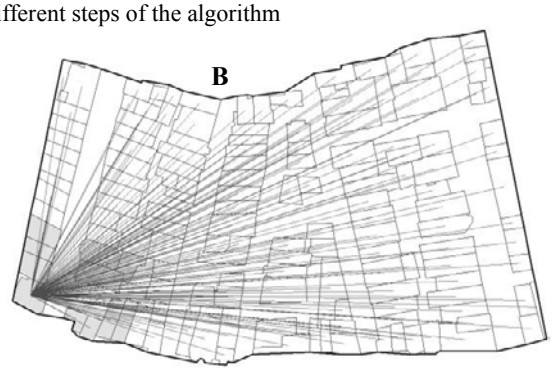
Figure 2: Starting feature selection algorithm

```
Data: Input FeatureSet G
Result: Feature f
    /* the function for getting the bounding line of a feature set
1   Function BoundLine (FeatureSet)
2       return BoundLine
    /* applying the function to the input feature set
3   BNDline ← BoundLine (G)
    /*declaring an empty feature array
4   AlongBNDlineFeatureArray ← ∅
    /* adding feature to the array that intersects the bounding line
5   for each Feature ∈ G do
6       if Feature ∩ BNDline do
7           Add Feature to the AlongBNDlineFeatureArray
    /*find the suitable corner feature
8   for each Feature1 ∈ AlongBNDlineFeatureArray do
        /* an array for holding the distance between a feature and all
        /*other features of the AlongBNDlineFeatureArray
9     FeatureDiastanceArray ← ∅
        /*calculating distances between the a feature and all other
        /* features of the AlongBNDlineFeatureArray
10      for each Feature2 ∈ AlongBNDlineFeatureArray do
11          Calculate distance between Feature1 and Feature2
12          Add the distance in the FeatureDiastanceArray
        /*finding the maximum distance for each feature by sorting
        /*the array descending and get the first element
        sort descending FeatureDiastanceArray (distance)
13      MAXVvalue ← first element of FeatureDiastanceArray
14      Tag the MAXValue to Feature1
    /*find the feature by sorting AlongBNDlineFeatureArray
    /*discending with regards to the tagged value and get the first
    /* feature of the array
15  sort descending AlongBNDlineFeatureArray (Tagged value)
16  return first element of AlongBNDlineFeatureArray
```

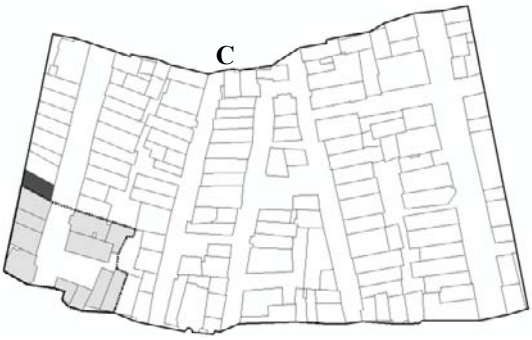
Figure 3: Visual illustration of different steps of the algorithm



The features with gray color touch the bounding line (black line). For each gray feature distance to other gray features are measured and the maximum is stored. The maximum of gray feature 1 (thicker line) is higher than the maximum of gray feature 2 (thicker line). Thus the gray feature whose maximum is the highest is selected as starting feature



A region (feature set with gray color) is formed by aggregating features from *G* with starting feature based on closest distance



A feature (dark gray) from *G* is selected as a start feature for subsequent region building if it is closest to the bounding line of (thick dotted line) previously formed region and touches the bounding line of the input feature set *G*

All the features along the bounding line (outer black line) of *G* are classified into regions. A second bounding line (inner black line) is formed for the unclassified features.

Step3. Formation of the first region: At the beginning the first region R_1 is formed only with the starting feature. Then the region is grown by grouping features from *G* on the basis of minimum distance, which means a feature from *G* is allowed to be grouped with the starting feature if the distance between them is a minimum compared to the distance of other features in *G* (fig. 3B). This grouping or region building is continued until $O(R_1) = T \pm d$ criteria is fulfilled. Since a feature in *G* is not allowed to divide according to the underlying data model, it is only possible to completely include or exclude a feature to a region. Which means the feature cannot be sliced. So, $O(R_1)$ cannot always be exactly equal to *T*. The maximum possible deviation of $O(R_1)$ with *T* for R_1 to R_{N-1} will be thus the maximum value of $|A|$ of any given *G*.

Once a region R_i is formed, a static variable *StatN* is updated with the number of region formed and the feature set on which the process will be continued is obtained by $G - R_i$. The process terminates and goes out of scope when $N\text{-}1 = StatN$. For example, if 3 regions are expected and 2 regions have already been completed then remaining features of *G* automatically form a region and the process goes out of scope. The figure 4 presents the algorithm of Step3.

Figure 4: Algorithm for first region formation

```
Data: Starting feature
Result: FeatureSet (Region), updated G
    /* static variable for keeping track of the regions
1   RegionFormed = 0
    /*an array of features representing a region
2   RegionFeatureArray ← ∅
3   Add starting feature to RegionFeatureArray
    /*declaring an empty feature array
4   FeatureArrayWithDistance ← ∅
    /* adding feature to the array with distance to start feature
5   for each Feature ∈ G do
6       Calculate Distance between Feature and Starting Feature
7       Tag Distance to Feature
8       Add Feature with distance to FeatureArrayWithDistance
    /*sorting the array of feature with regard to the distance
9   sort ascending FeatureArrayWithDistance (distance)
    /*adding closest features to the region array until region's
    /* O(Ri) = T ± d
10  for each Feature ∈ FeatureArrayWithDistance do
11      if ∑|A| of RegionFeatureArray < T do
12          Add Feature to RegionFeatureArray
13      else do
14          finalize RegionFeatureArray
15          RegionFormed = RegionFormed +1
    /*returning the region and updated G
16  return RegionFeatureArray
17  return G ← G - RegionFeatureArray
```

Step4. Start feature selection for subsequent regions: As stated earlier, the algorithm prioritizes forming regions along the bounding line *BNDline(G)* of the input feature set *G*. Therefore, a start feature for any subsequent region R_{i+1} should be located next to the former region R_i and also should touch the *BNDline(G)* (fig. 3C). These are two simple criteria for selecting a start feature for any subsequent region building. The algorithm of this step is given in next page with figure 5.

Figure 5: Algorithm for start feature for subsequent regions

```
Data: Feature Set Ri (last region), updated G
Result: Feature f
    /* the function for getting the bounding line of a feature set
1   Function BoundLine (FeatureSet)
2       return BoundLine
    /* applying the function to the input region feature set gi
3   BNDline_Ri ← BoundLine (Ri)
    /* applying the function to the feature set G
4   BNDline_G ← BoundLine (G)
    /*declaring an empty feature array
5   FeatureSet_near_ BNDline_Ri ← ∅
    /* adding feature from G to the array that are with a certain
    /* distance (5m) from the BNDline_Ri
6   for each Feature ∈ G do
7       if Feature is within 5m of BNDline_Ri do
8           Calculate distance from Feature to BNDline_Ri
9           Tag distance to Feature
10          Add Feature to the FeatureSet_near_ BNDline_Ri
    /*sort FeatureSet_near_ BNDline_Ri in ascending way so that
    /* the feature with shortest distance comes first in a loop
11  sort ascending FeatureSet_near_ BNDline_Ri(distance)
    /*find the start feature
12  for each Feature ∈ FeatureSet_near_ BNDline_Ri do
13      if Feature touches BNDline_G do
14          return Feature
```

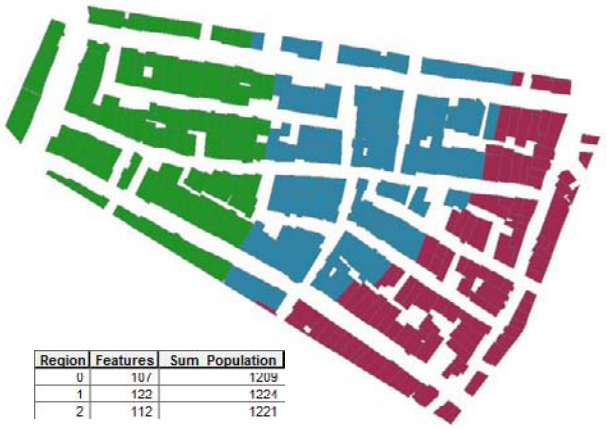
Step5. Repetition: Step 3 to 4 are repeated until no start feature is returned by step 4 and the required number of regions is still not achieved. A null feature return by step 4 means all the features along the bounding line of *G* are classified into regions. If this is the case, a new bounding line is created for the set of non-classified features (fig. 3D). The *BNDline(G)* which is created in step2 is replaced by the new bounding line and the process starts continuing from step 2.

4 Implementation, application and results

The algorithm we presented in section 3 has been implemented using c# programming language and ArcObjects library of ESRI. Figure 6 and 7 shows the result of 2 examples of an application of the implemented algorithm. Each feature (polygon) in both figures represents residential buildings and has an attribute called population (no. of residents). The maximum value of *d* of the input feature set was 21.

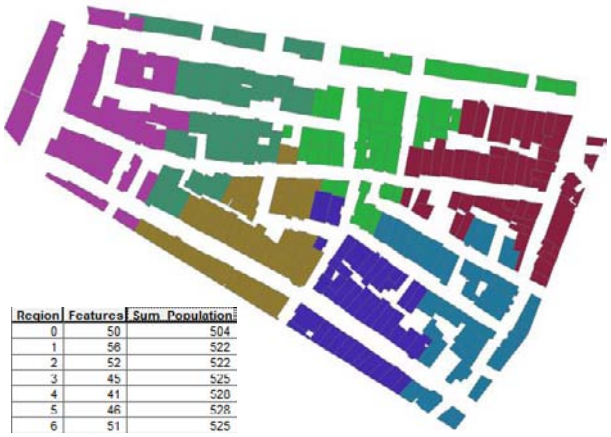
In figure 6, the expected number of equitable regions was 3 based on the population attribute which means the feature set has to be divided into 3 non-overlapping regions so that the total population for each region remains approximately equal. In figure 7 the expected region number was 7. Both figures show a distinct division of the feature set into regions. None of the region in both figure overlap with others. However, the shape of the regions gets more irregular with the increase number of regions (fig.7). The important point to be noted here is that region no.0 in both figures differs significantly from other regions in terms of total population and the difference goes beyond the *MAX(d)* in figure 7. The differences among other regions are minimal and within *MAX(d)*.

Figure 6: Input feature set *G* divided into 3 equitable regions



Region	Features	Sum_Population
0	107	1209
1	122	1224
2	112	1221

Figure 7: Input feature set *G* divided into 7 equitable regions



Region	Features	Sum_Population
0	50	504
1	56	522
2	52	522
3	45	525
4	41	520
5	46	528
6	51	525

Region 0 is in fact the last region formed with the remaining feature set once *N-1* regions are formed. If the other regions formed with a positive value of *d* (section 3, step 3) then the effect goes on to the last N^{th} region (region 0) which is forced to be formed with a total value deduced by the cumulative positive *d* of the former regions. Thus only the N^{th} region's $O(R_n)$ may not be equal to $T \pm d$. The maximum difference between the last region's $O(R_n)$ with other regions $O(R_i)$ is thus expected to be higher with the increased no. of regions. However, this problem can be solved with a constraint that two consecutive regions should form with $+d$ and $-d$ simultaneously which restricts region formation with always $+d$ or $-d$.

Another important point to be noted here that theoretically, a feature set may contain several features which are suitable as starting feature for the formation of first region (step 2, section 3). Selection of each of those suitable features as a start feature would results a different form of the output regions in terms of region's $O(R_i)$ and shape. The algorithm

restricts different output possibilities by automatically selecting the best starting feature. But the algorithm could be adopted for allowing the selection of alternative suitable features for obtaining variations in output regions.

5 Conclusion and future works

The paper presented an algorithm for segmenting a feature set into multiple equitable non-overlapping regions and its implementation and the result of its application are discussed. As a first attempt the algorithm has been developed and implemented to deal with polygon and point features set. Due to the limited space examples of point feature sets are not discussed in this paper. Several tests have proven its applicability. The applicability could be extended to polyline features set in future with limited effort. At present the algorithm is dealing with a single attribute and regions are formed based on a value which is approximately equal for each region. As a future work the algorithm could be extended to deal with multiple attributes and other statistical parameters e.g. regions could be formed based on equal standard deviation of one or multiple attributes. The algorithm could be enriched by introducing constraints e.g. region formation can be restricted to cross certain types of roads and other geographical features. Moreover, spatial indexing could be applied to improve computing time for large input datasets.

References

[1] ESRI. White paper on territory design. Redlands, USA, 2010.

[2] N. Shortt. Regionalization/zoning systems. In R. Kitchin and N. Thrift, editors, *International Encyclopedia of Human Geography*, pages 298--301. Elsevier, Oxford, 2009.

[3] S. Alvanides, S. Openshaw and P. Rees. Designing your own geographies. In P. Rees, D. Martin and P. Williamson, editors, *The Census Data System*, pages 47—65. JohnWiley, Chichester, Sussex, 2002.

[4] S. Alvanides. Zone Design Methods for Application in Human Geography. *PhD thesis*, School of Geography, University of Leeds, 2000.

[5] S. Cockings, A. Harfoot, D. Martin and D. Hornby. Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 Census output geographies for England and Wales. *Journal of Environment and Planning*, A 2011, volume 43: 2399 - 2418, 2011.

[6] S. Openshaw. *A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling*. Transactions of the Institute of British Geographers, New Series 2, 1977.

[7] S. Openshaw. Algorithm 3: a procedure to generate pseudo-random aggregations of N zones into M zones, where M is less than N. *Journal of Environment and Planning*, A 9: 1423-1428, 1977.

[8] S. Openshaw and L. Rao. Algorithms for re-engineering 1991 Census geography. *Journal of Environment and Planning*, A 27: 425 - 446, 1995.

Session:
Analysis and Education

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Geographic Information Technologies for analysing the digital footprint of tourists

Toni Hernández, Rosa Olivella, Josep Sitjar, Lluís Vicens
University of Girona – SIGTE
Pl. Ferrater Mora, 1
17071 Girona (Spain)
info@sigte.udg.edu

Abstract

As part of a study on the use of the city by visitors, this paper discusses the technical solution adopted in order to analyse and exploit the geo-digital footprint generated by the said visitors using GPS devices.

1 Background

For the purpose of analysing the use of the city by tourists, the tourism research group (INSETUR) of the University of Girona planned a project in the city of Girona that involved the collection of movement data and its subsequent analysis using geospatial techniques applied by the SIGTE (the GIS Centre of the University of Girona), the results of which are presented in this document.

The methodology used by the research group to collect this data was a GPS device provided by the Tourist Office that visitors had to carry on their person throughout the one-day visit of the city. At the end of the visit, the tourists returned the device and answered a qualitative questionnaire.

The project has collected tourist movement data over the course of 9 months, generating a total of 1,339 tracks or 4,752,804 waypoints and 1,339 questionnaires. This paper describes the GIS technical solution that has been adopted in order to carry out the geospatial analysis of this data.

2 Introduction

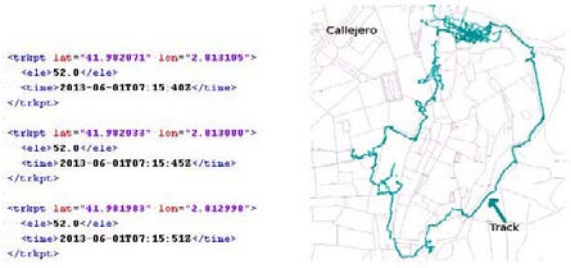
In order to fulfil the stated purpose, the research group carried out an analysis to identify the city streets most visited by tourists and the capacity of attraction of certain heritage elements.

The term "city" in this case refers to the historical centre, clearly delimited by natural elements (river) or anthropic elements (walls), which has been established as the area of study.

In order to carry out this study, over 1,339 GPS tracks have been captured thanks to the collaboration of tourists. These files (in .gpx format) contain, with a certain margin of error, all the information on the itinerary completed by the visitors to the city (route, distribution of time along the route, etc.) and form the basis of this project.

A track is an ordered list of points. The coordinates (latitude and longitude) of each point are known, along with the date and time when each coordinate was captured.

The image below shows a fragment of the track file in gpx format. The contents of the track file are displayed on the left and the graphic representation of the track on the city street map is displayed on the right.



3 Methodology

3.1 Tools

Having evaluated several work tools (both proprietary and open source), the research team has opted to import the files in .gpx format into a **PostgreSQL/PostGIS database** and to

use **PostGIS spatial tools** to analyse the information contained in the .gpx files on the basis of **SQL sentences**.

3.2 Technical challenges

The main challenge of the project consists of breaking down each track into a list of streets (or street arrays) along which each track passes. As such, it will be possible to know the sections of the streets through which each track passes and the length of time invested in each of these sections.

The area of study itself poses a challenge, since the typical dim, narrow streets of Girona old quarter are susceptible to poor satellite coverage [1]. It is therefore necessary to correct the tracks generated in order not to lose too much data for the subsequent analysis.

This challenge will consist of structuring the data in such a way that the planned analysis is both possible and agile.

All of the processes carried out are listed below.

3.3 Importing the GPS data

In order to import the .gpx files, the ogr2ogr tool has been used [2]. From a command console the following command is keyed in to import the .gpx track file into the track table within the TURISMO database.

ogr2ogr -f "PostgreSQL" PG:"host=localhost user=postgres port=5433 dbname=TURISMO schemas=originals password=contraseña" track.gpx -overwrite -lco GEOMETRY_NAME=geom track_points -nln "track"

This command uses the Postgres username and password to connect to the TURISMO database. Once connected it generates a new track table with all the information contained in the .gpx track file.

gid integer	time character varying(254)	lat double precision	long double precision	fecha character varying(254)	hora character varying(254)	geom geometry(Point)
1	2013/03/13 9:41:00	485399.529832	4648107.63786	2013/03/13	9:41:00+00	01010000020F759
2	2013/03/13 9:41:02	485398.275501	4648109.36234	2013/03/13	9:41:02+00	01010000020F759
3	2013/03/13 9:41:03	485392.042009	4648147.06035	2013/03/13	9:41:03+00	01010000020F759
4	2013/03/13 9:41:04	485392.452281	4648144.58313	2013/03/13	9:41:04+00	01010000020F759
5	2013/03/13 9:41:05	485391.753827	4648140.47258	2013/03/13	9:41:05+00	01010000020F759
6	2013/03/13 9:41:06	485390.931959	4648138.78613	2013/03/13	9:41:06+00	01010000020F759
7	2013/03/13 9:41:00	485390.044762	4648137.00044	2013/03/13	9:41:00+00	01010000020F759

The image above shows a fragment of the contents of the track table after it has been imported. The *gid* field functions as the primary key and serves to identify unequivocally each point in the table. This attribute will subsequently be used in new SQL sentences. The geometry of the points is found in the *geo* column, while the *time* column contains both the date and the time at which each track point was captured.

The bulk import of the .gpx files, of which there are over 1,600, is carried out through a command file (.bat). This command file is generated from a script that reads the contents of a folder (where the tracks are located) and generates an import sentence for each .gpx file.

Once each track is imported it is necessary to convert it from geographic coordinates (GPS) to projected coordinates (UTM). This step is obligatory in order to carry out a subsequent Snap operation indicating a tolerance in metric units (corresponding to the projected coordinates).

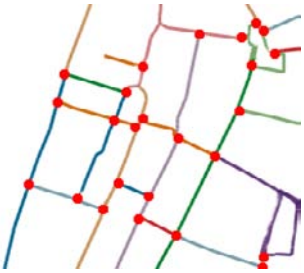
With the following SQL sentence a new table is created with all the initial columns, plus the geometry contents re-projected to the ETRS89 coordinate reference system (srid = 25831).

create table track_reproyectado as
select st_transform(geom,25831) as geom_reproyectada, track. from track;*

3.4 Street map

In order to carry out this operation it is necessary to import a street map of the area of study into the database. The main characteristic of this street map is that every street is split into various sections. Each section of each street is represented by a specific spatial entity. In other words, the number of sections into which each street is split is equal to the number of intersections that this street has with other streets. By splitting the streets in this way it is possible to detect the specific street section along which each track passes, while at the same time distinguishing the sections of the same street along which the track does not pass.

The image below displays the points of intersection between streets (marked in red). The various sections of the same street are displayed in a single colour.



The shp2pgsql tool has been used [3] (included in PostGIS) to import this shp format street map into the database.

From a command console we execute:

Shp2pgsql callejero_shape tabla_callejero > callejero.sql

This command will generate the callejero.sql file with the contents of the .shp file translated into SQL expressions. The next step consists of executing the callejero.sql file within the TURISMO database.

The file can either be executed from pgAdmin or, from a command console, by using the PostgreSQL command line interactive client, called psql, as shown below.

psql -d TURISMO -U postgres -W -f callejero.sql

This command will request the password of the corresponding user (-U).

3.5 Assigning the track points on the street map

Each point located on the track of the GPS device must be assigned to the nearest street. As such we can generate a list with the names of the streets along which the track passes. Only the track points located no more than 7 metres from a street have been assigned; the remaining track points have been discarded.

In order to assign each track point to a street the following SQL command has been executed, inspired by Paul Ramsey's post [4]:

```
create table track_callejero as
      select distinct on (punto_id) c.gid as calle_id, tp.gid as punto_id
         from track_reproyectado tp
         inner join callejero c on st_DWithin(tp.geom, c.geom, 7.0)
      order by tp.gid, st_Distance(c.geom, tp.geom);
```

This command calculates, for each track point, which streets are less than seven metres away. If a track point has more than one street located less than seven metres away, it places the streets in order so that the nearest street appears in first place. Finally, the distinct (*punto_id*) clause indicates that we only wish to obtain one street for each point. The fact that the streets have previously been placed in order means that each track point is assigned exclusively to its nearest street.

The result of the sentence above is a new *track_callejero* table that lists each track point together with the nearest street. The matching of track points and streets is carried out on the basis of the alphanumeric attributes (gid) that identify each point and each street. These 'gid' attributes are renamed *punto_id* and *calle_id*, respectively, in order to facilitate their interpretation.

calle_id integer	punto_id integer
339	1
350	5
350	6
350	7
350	8
350	9
350	10

The streets from the above list for which fewer than five track points have captured have been omitted from the final result. As such we have omitted certain sections that are not very representative of the itinerary followed by the track.

3.6 Time variable

Having obtained the streets along which the track passes we can also find out the order in which the streets have been visited and the amount of time invested in each street.

In order to extract this information it is necessary to create a PL/PSQL function called 'itinerary'. The result returned by this function is an array data structure. Each element of the array contains three variables: *calle_id*, *punto_id* and *contador*, which is the variable used to count the number of times the same track passes along the same street.

```
CREATE OR REPLACE FUNCTION itinerario(_tbl character varying, _col1 character varying, _col2 character varying)
RETURNS SETOF integer[] AS $$
DECLARE
      ultima_calle integer;
      contador integer;
      fila RECORD;
      r boolean;
BEGIN
FOR fila IN EXECUTE 'SELECT ' || quote_ident(_col1) || ' as calle_id, ' || quote_ident(_col2) || ' as punto_id FROM ' || quote_ident(_tbl) || ' ORDER BY ' || quote_ident(_col2)
      LOOP
      IF ultima_calle IS NULL THEN
         ultima_calle:= fila.calle_id;
         contador:= 0;
      ELSE
         IF ultima_calle != fila.calle_id THEN
               ultima_calle:= fila.calle_id;
               contador:=contador+1;
         END IF;
      END IF;
       RETURN next ARRAY[ fila.calle_id, fila.punto_id, contador];
END LOOP;
END
   $$ LANGUAGE plpgsql VOLATILE;
```

As can be observed, the itinerary function receives three parameters: the name of the table and the names of the columns that contain the identifiers of the track points and the streets. These three parameters correspond to the previously created *track_callejero* table.

With the data structure returned by the itinerary function we can obtain a list of the first and last track points that pass along each street. For this purpose we execute the following command:

```
create table itinerario_seguido as
      SELECT calle_id, MIN(punto_id) as primer_pt_id, MAX(punto_id) as ultimo_pt_id FROM (
            SELECT ar[1] calle_id, ar[2] punto_id, ar[3] contador from (select itinerario('track_callejero','calle_id','punto_id') ar) as fo
         ) as foo
      GROUP BY calle_id, contador
      ORDER BY contador;
```

We obtain a new list, as can be observed in the image below.

calle_id integer	primer_pt_id integer	ultimo_pt_id integer
339	1	1
350	5	39
1636	40	42
352	43	65
2561	66	70
2560	71	76

With this new table we can then calculate the time invested in each street, using the time that has elapsed between the first and the last point of each street.

Select calle_id, (t2.time – t1.time) as tiempo_invertido_en_calle
from track t1, track t2, itinerario_seguido i
where i.primer_pt_id= t1.gid and i.ultimo_pt_id = t2.gid;

calle id integer	tiempo invertido en calle interval
350	00:00:52
1636	00:00:02
352	00:00:39
2561	00:00:05
2560	00:00:07
1369	00:00:13
2557	00:00:03
2468	00:00:03
198	00:00:01

3.7 Assigning track points to heritage elements

Certain heritage elements of the city are defined as key attractions by the city's tourism promoters. As such, we have set out to analyse the time that visitors (object of the study) devote to each of them.

It is considered that a heritage element has been visited when the tourist has remained at the site for a predefined length of time (15 minutes in the case of elements for which a short visit time is considered sufficient, and 30 minutes in the case of those for which more time is considered necessary).

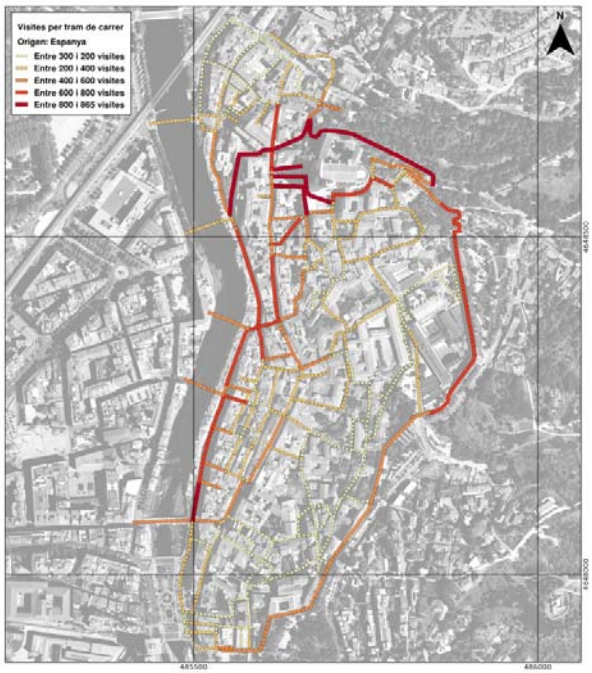
An area of influence has been generated around these heritage elements, and the first and last track points detected in this area of influence have been identified. Given that each track point contains information on the hour/minute/second in which it has been captured, it is straightforward to calculate the length of time devoted by the tourist to visiting each element.

Linear heritage elements, such as the city walls and certain streets, are considered visited when the track points closely follow them.

4 Results

The possibility of comparing each visitor's track with a qualitative questionnaire has made it possible to obtain specific results in relation to some of the questions that were posed. As such it has been possible to represent the assignment of tracks on the street map (that is, represent the degree of intensity of use of the street map), not only for all the visitors who formed part of the study but also according to the following factors:

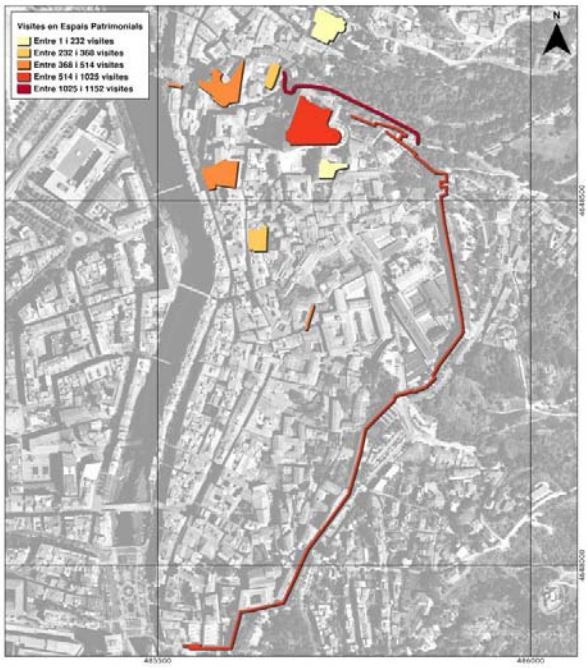
- Where they have stayed (accommodation)
- Nationality
- Number of times they have visited the city previously
- Reason for the visit



It has also been possible to represent other aspects thanks to the time data contained by the tracks, such as specific times of day or seasons of the year.

As in the case of streets (or street arrays), the visits to key heritage elements have also been considered in their totality, as well as in relation to the different variables gathered from the completed questionnaire. Thanks to the design and implementation of a relational database that contains all the information on the street arrays, questionnaire answers, heritage sites and tourist tracks, it is possible to produce specific theme-based cartography by combining the information provided by all of these elements.

Furthermore, a representation is also displayed of the track points prior to being assigned to the street map, in order to capture the total magnitude and for the purpose of comparing it with the 'manipulated' results after assigning the points to the street map.



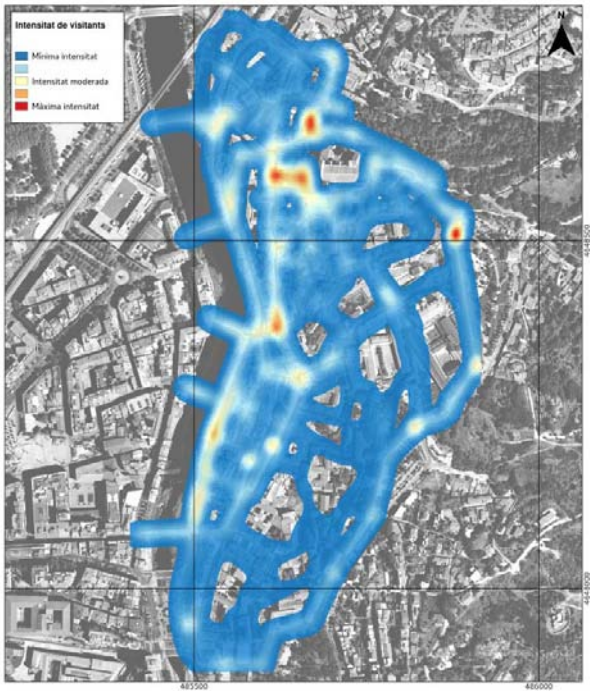
The result of this general analysis of waypoints has been displayed through a density map [5]. This type of technique for the interpolation of point data compares the space with the number and position of each of the waypoints for the total area of study. A density map, which to a large extent offers qualitative results, in this case provides information on which areas or spaces of the city are most densely occupied by tourists, while at the same time offering information on where in the city it would be easiest to find the highest concentration of tourists at a given time of day or in a given season of the year.

5 Conclusions

The project is an example of how databases can be used to carry out spatial analyses. In this respect it should be taken into account that the entire project, except for the final graphic representation, has been carried out within the PostgreSQL/PostGIS spatial database using SQL (Structured Query Language) commands.

Through these SQL commands we have carried out the following operations:

- The street map has been broken down into a segmented street map in order to analyse which street sections are the most visited.
- The GPS tracks have been converted from geographic coordinates to projected coordinates. In this way we have been able to use Euclidean distances to assign each track point to the nearest street.
- An analysis has been carried out of the most visited areas of the city, taking into account specific times of day and seasons of the year.
- The results obtained have been compared with the results of the questionnaires completed by the tourists. As such it is possible to know what tourists' preferences are according to a certain parameter of the questionnaire (nationality, budget, etc.).



Without a doubt, this project is a successful case of a spatial analysis carried out not using a conventional desktop GIS (the most habitual solution) but rather a robust open source database (Postgresql) supplemented with a highly developed and stable spatial capability (POSTGIS).

This opens up a range of possibilities for working on and exploring other functions of analysis or visualisation that continue to add value to field-gathered data.

References

[1] M. Modsching, R.Kramer and K.ten Hagen. Field trial on GPS Accuracy in a medium size city: The influence of built-up in Proceedings of the 3rd Workshop on Positioning, Navigation and Communication. 2006

[2] http://www.gdal.org/ogr2ogr.html

[3] http://postgis.net/docs/manual-1.3/ch04.html, ch. 4.3.2 & http://manpages.ubuntu.com/manpages/natty/man1/shp2pgsql.1.html

[4] http://blog.cleverelephant.ca/2008/04/snapping-points-in-postgis.html

[5] V. Olaya, Sistemas de Información Geográfica (2013)

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

"Troy is ours – How on earth could Clytaemnestra know so fast?"

Titus Tienaah and Emmanuel Stefanakis
University of New Brunswick, Department of Geodesy and Geomatics Engineering
P.O.Box 4400, Fredericton, NB, E3B5A3, Canada
{ ttienaah, estef }@unb.ca

Abstract

This paper introduces a web application developed to highlight Aeschylus' perception on the transmission of a message from Troy to Mycenae through a series of beacons; a case study that investigates the veracity of an ancient spatial description to facilitate message transmission over a distance of approximately 500 km. This research demonstrates the use of modern web and geospatial tools to recreate and animate a series of events of the ancient world through geo-visualization. The outcome of this work serves as an educational resource to supplement mythological text and stories passed down by oral poetry and storytelling.
Keywords: geospatial web, animated maps, 3D visualizations.

1 Introduction

This paper presents an application developed using geospatial web tools to visualize the transmission of a very important message in the ancient world. The message travelled a distance of approximately 500 km in just a few hours. Agamemnon, king of Mycenae, along with other Greek kings led an expedition of Achaean troops to Troy and besieged the city for ten years, until the night the city fell to the ruse of the Trojan Horse. The following morning, Agamemnon's wife, Clytaemnestra, in Mycenae was already aware of the glory (Figure 1).

Figure 1: Troy and Mycenae.

Source: Google Earth.

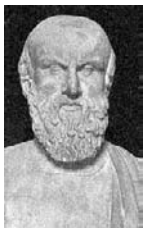
Historical event of 12th century BC or just another story of the Greek Mythology, the Trojan War has been narrated through many works of Greek literature, most notably through Homer's Iliad and the Odyssey. The events and details of the Trojan War were passed on orally in poetry, non-poetic storytelling, and vase painting through the centuries. The great tragic playwrights of Athens, Aeschylus, Sophocles, and Euripides, were infused by those myths and wrote a series of relevant ancient dramas.

Their tragedies are extremely valuable resources of the ancient world as they help us better understand the social structure, political system, faith, beliefs, and even the scientific knowledge and technological achievements of their era.

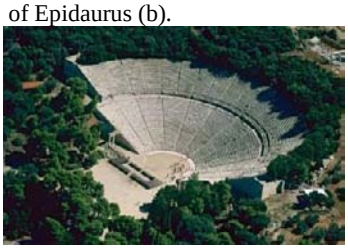
This work was inspired by an act from Agamemnon tragedy. The main objective of this case study is to investigate the likelihood of intervisibility between beacons from Troy to Mycenae. To answer the question in the title of this paper, additional questions were posed: how can intervisibilty between beacons be visualized in two and three dimensions; and how this possibility can be tested using modern web and geospatial tools?

A web map application has been developed to highlight the transmission of the message "Troy is ours" from the city of Hector and Paris to the kingdom of Mycenae, using a series of synchronized beacons. This is how Aeschylus (Figure 2a) describes in 25 lines the tele-communication means of the 12th century BC. It is stunning how accurate the geography of the Aegean is unfolded through those lines. After attending the play of the tragedy in the ancient theatre of Epidaurus (Figure 2b) in summer 2013 (Agamemnon, 2013), we examined closely those lines (Fagles, 1977), and thought that this transmission should be visualized on maps!

Figure 2: Aeschylus 525-456 BC (a), and the Ancient Theatre of Epidaurus (b).



(a)



(b)

Source: Wikipedia.

Figure 3: The transmission of victory message in yellow line. KML/Z file is available at: http://gaia.gge.unb.ca/troyisours/KML)

Source: Google Earth.

The following Sections present a case study that clearly demonstrates the potential of geospatial web tools in the study of the ancient world through efficient educational tools. Section 2 provides a short description of the act and a map of beacons and places mentioned by Aeschylus. Section 3 presents the tools and functionality of the web application. Section 4 highlights the outcomes of this project.

2 Transmission of the Message

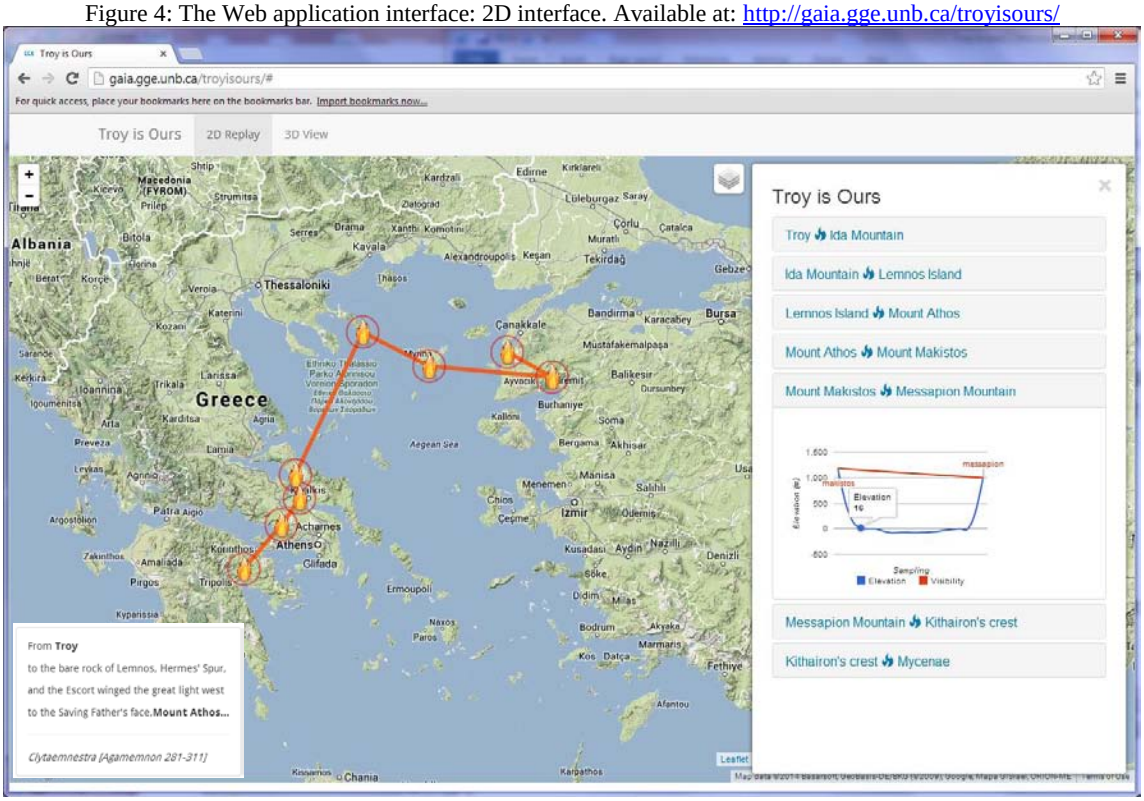
Agamemnon is the first play of the only complete trilogy that has come down from antiquity, entitled the Oresteia (Fagles, 1977) and written by Aeschylus in 458 BC. In this play, Aeschylus describes the return of king Agamemnon from his victory in the Trojan War, from the perspective of both the town's people (the Chorus) and his wife, Clytaemnestra. The story unfolds to the death of the king at the hands of his wife. Clytaemnestra was angry at his sacrifice of their daughter Iphigenia, killed so the gods would stop a storm hindering the Greek fleet in the war. She was also unhappy at his keeping of the Trojan prophetess Cassandra as a concubine. The ending of the play includes a prediction of the return of Orestes, son of Agamemnon, who will seek to avenge his father.

In the beginning of the play (lines 264-312), Clytaemnestra and the Chorus are on the scene. Clytaemnestra proudly announces that "the Greeks have taken Troy!" As "…the joy is beyond [the Chorus] hopes…", the Leader of the Chorus keeps asking the Queen "…and you have proof?" and "…when did they [Achaeans] storm the city [Troy]?" Then, Clytaemnestra replies: "Last night, I say, the mother of this morning." It is when the Chorus' Leader raises his voice to doubt the news. He says: "And who on earth could run the news so fast?"

"Hephaestus, the god of fire… and beacon to beacon rushed it [the message] on to me", Clytaemnestra will respond. In the next 30 lines, Aeschylus, through the voice of Clytaemnestra, will take the audience and fly with them across the Aegean Sea from Troy to the "bare rock of Lemnos" and the Mount Athos. From there, over Euboea and the Saronic Gulf to Argolis, where the Mycenae Palace is located. Aeschylus' description is so poetic and fruitful, including details about the geomorphology, history, fauna, and flora of the Achaean land.

Figure 3 provides a visualization of the beacon locations in Google Earth. The yellow line represents the visual passing of the message from Troy to Mycenae through the beacons. The images around the map are perspective views generated in Google Earth. The terrain (elevations) was exaggerated by a factor of 3 to highlight the Earth's relief. By examining the perspective images, it is clear that the visibility between the beacons described Aeschylus is valid. Hence, Aeschylus description makes sense. The total distance from Troy to

Figure 4: The Web application interface: 2D interface. Available at: http://gaia.gge.unb.ca/troyisours/



Mycenae through the beacons equals to 540 km. The longest segment is the one connecting Mount Athos to Mount Makistos which is equal to 180 km.

Obviously, Aeschylus had a very clear understanding of the Earth's relief. In addition, these 30 lines convey lots of details about the geography of the area. In this study, not much attention is spent on those details; we are planning to investigate them closely in collaboration with archaeologists in the future.

The next Section introduces a web application that was designed and implemented as part of this project. The application makes use of modern geospatial web tools. Both the tools and functionality are described. The main scope of the application is to serve as an educational tool to both school students and the general public.

3 Web Application

The geospatial web application uses multiple modern web technologies: HyperText Markup Language (HTML 5), Cascading Style Sheets (CSS 3), Scalable Vector Graphics (SVG) and JavaScript. To achieve responsiveness across multiple devices (e.g. mobile, desktop and tablet browsers), the application uses Bootstrap 3 (2014) as a front-end framework. Two dimensional interactive maps are created using Leaflet (2014) JavaScript library. Through Leaflet, SVG and other vector geometry are overlaid on top of two tile layers: Google Maps (Terrain) and Open Street Map (OSM). SVG graphics and animation are enabled through the use of Data Driven Documents (D3, 2014) JavaScript library. Google Earth application programming interface (GE API, 2014) is used for 3D visualization and tour animation. Other dependant JavaScript libraries include jQuery (2014) and Lo-Dash (2014).

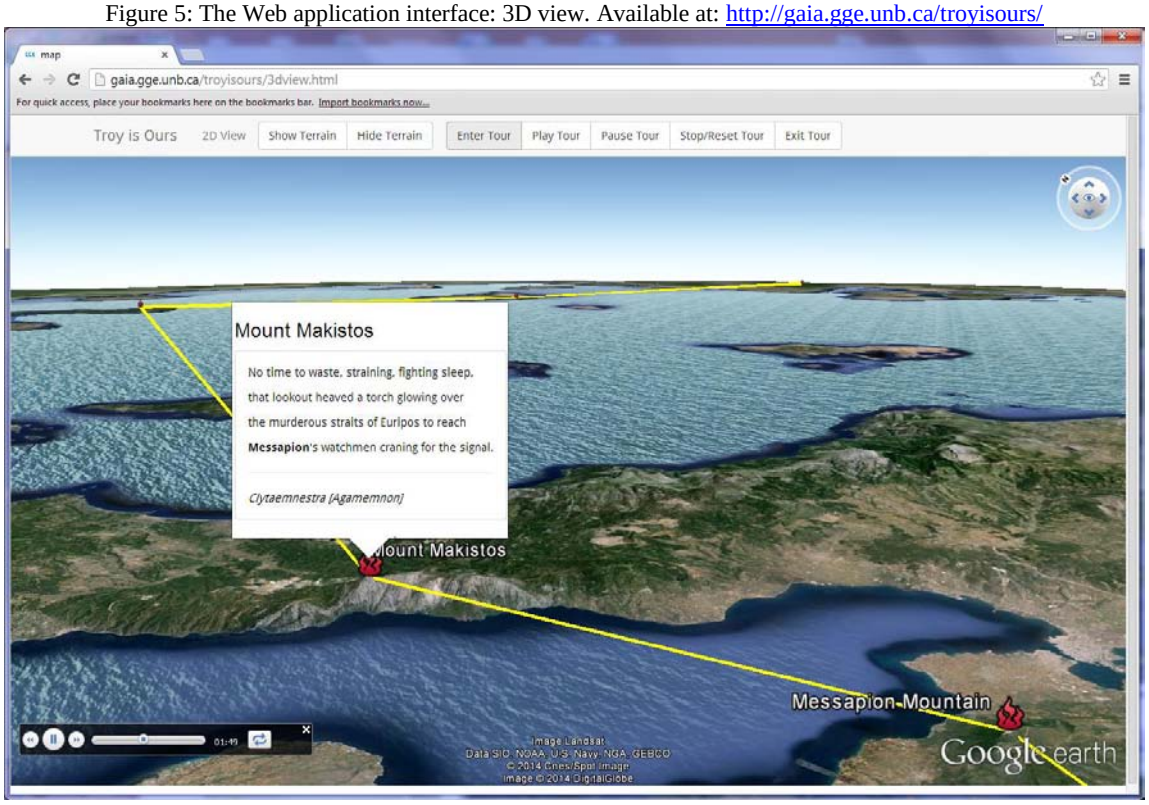
The application has two functional interfaces: 2D visualization and animation and 3D visualization and touring. The 2D interface (Figure 4) uses map tiles as base layers (e.g., Google Maps – Terrain) with SVG and vector animations to describe transmission of the message "Troy is ours". At each chain of the transmission (from one beacon to the next) Google Elevation Service (2014) is queried for terrain elevations along the chain. Using Google Visualization API (2014), a profile is visualized as line chart. Since each signal is transmitted using fire, a straight line plot from the origin to destination is used to represent light communication. If the 2D profile does not intersect the light transmission between the origin and destination, then indivisibility occurs between these two beacons.

The 3D view uses Google Earth browser plugin (Figure 5). The application uses a pre-recorded tour stored in KML to visit each station (from Troy to Mycenae) with terrain visualization and annotated text from Aeschylus. A user can enter the tour and navigate from beacon to beacon following Aeschylus description.

4 Discussion

Examining 2D profiles and perspective 3D views, it is clear that the visibility between the beacons described by Aeschylus is valid. Although there are lots of room for interpretation of

Figure 5: The Web application interface: 3D view. Available at: http://gaia.gge.unb.ca/troyisours/



Aeschylus play by experts (such as archaeologists, historians, and educators), we aspire that this case study and the web application can serve as a research and educational tool to test and explore various scenarios. Our future plans include a close collaboration with two main objectives: (a) to enrich the application by including their perspective (e.g., alternative opinions and related stories); and (b) build and test an educational program for school students.

A similar approach and collaboration with the Educational Programs Department in the Archaeological Museum of Athens has led in the past to the development of a web application (Stefanakis, 2012, 2013) for the origin of the Antikythera mechanism (2013), the world's oldest known analogue computer, which is exhibited at the Museum. The application was approved and included in the educational program of the Museum. Since then, it has been introduced to hundreds of school students and visitors, and has received very positive comments.

References

Agamemnon, 2013. Municipal Theatre Kozanis, Athens and Epidaurus Festival, 2013.

Antikythera Mechanism, 2013. Antikythera Mechanism Research Project. http://www.antikythera-mechanism.gr/ [last visited: March 1, 2014]

Bootstrap, 2014. A Front-end Framework for Developing Responsive, Mobile First Projects on the Web. http://getbootstrap.com/ [last visited: March 1, 2014]

D3, 2014. Data Driven Documents. http://d3js.org/ [last visited: March 1, 2014]

Fagles, R., 1977. Aeschylus: The Oresteia (Agamemnon, The Libation Bearers, The Eumenides). Translation. Penguin Classics.

GE API, 2014. Google Earth Application Programming Interface. https://developers.google.com/earth/ [last visited: March 1, 2014]

Google Elevation Service, 2014. The Google Elevation API. https://developers.google.com/maps/documentation/elevation/ [last visited: March 1, 2014]

Google Visualization API, 2014. Showing Elevation along a Path. https://developers.google.com/maps/documentation/javascript/examples/elevation-paths [last visited: March 1, 2014]

jQuery, 2014. A Fast, Small, and Feature-rich JavaScript library. http://jquery.com/ [last visited: March 1, 2014]

Leaflet, 2014. An Open-Source JavaScript Library for Mobile-Friendly Interactive Maps. http://leafletjs.com/ [last visited: March 1, 2014]

Lo-Dash , 2014. A Utility Library Delivering Consistency, Customization, Performance, and Extras. http://lodash.com/ [last visited: March 1, 2014]

Stefanakis, E., 2012. Map Mashups and APIs in Education. In: Peterson, M. (Ed.). Online Maps with APIs and Mapservices. Springer.

Stefanakis, E., 2013. Map-mashups in the study of cultural heritage. In the Proceedings of the 2nd Workshop on Integrating 4D, GIS and Cultural Heritage. Leuven, Belgium.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Workforce Demand Assessment to Shape Future GI-Education – First Results of a Survey

Barbara Hofer
University of Salzburg/ Department of Geoinformatics – Z_GIS, Schillerstr. 30, 5020 Salzburg, Austria
barbara.hofer@sbg.ac.at

Gudrun Wallentin
University of Salzburg/ Department of Geoinformatics – Z_GIS, Hellbrunnerstr. 34, 5020 Salzburg, Austria
gudrun.wallentin@sbg.ac.at

Christoph Traun
University of Salzburg/ Department of Geoinformatics – Z_GIS, Hellbrunnerstr. 34, 5020 Salzburg, Austria
christoph.traun@sbg.ac.at

Josef Strobl
University of Salzburg/ Department of Geoinformatics – Z_GIS, Hellbrunnerstr. 34, 5020 Salzburg, Austria
josef.strobl@sbg.ac.at

Abstract

Geographic Information Science & Technology (GIS&T) is constantly evolving in scientific and technological terms. In 2006 the GIS&T Body of Knowledge (BoK) initiative has provided a domain inventory that serves as a structured basis for curriculum development. The content and structure of the BoK are currently undergoing revision. One of the projects addressing an update of the BoK is the project Geographic Information: Need to Know. In this project an assessment of current and future workforce demand and educational supply in the geographic information (GI) domain provide the basis for revising the BoK. This article reports on first results from a survey regarding GI workforce demand in Europe. People working in the GIS&T domain were asked to rate BoK knowledge areas related to their relevance in a professional working context. These ratings are differentiated by types of organizations and educational backgrounds of respondents. The report is rounded off with an outlook to the results on future competences identified by respondents.

Keywords: Geographic Information Science and Technology, Body of Knowledge, education.

1 Introduction

An inventory of key topics in a domain can provide the basis for composing educational programmes. A prerequisite is that the inventory is kept up-to-date. In the Geographic Information Science & Technology (GIS&T) domain, such an inventory is the Body of Knowledge (BoK). This article presents first results from a survey that aimed at evaluating the current fit between BoK knowledge areas and professional tasks of the GI workforce.

As the domain of Geographic Information Science and Systems has matured over the last decades, its educational foundation has also evolved. Under the lead of David DiBiase the University Consortium for Geographic Information Science (UCGIS) developed the GIS&T BoK [1]. This UCGIS initiative was the first comprehensive attempt to provide a domain inventory in a strictly hierarchical list of knowledge areas, units, topics and related learning objectives. The intention of the GIS&T BoK initiative was to provide a comprehensive and structured basis for curriculum development. The BoK aimed at allowing the design of adaptable curricula that define individualised pathways through its 1,660 educational objectives [2]. Further uses were expected to closely link to the geospatial industry, including programme accreditation, professional certification and the design of job descriptions. However, although the GIS&T BoK has been a milestone achievement and still is the main reference document for the geospatial domain, the document is largely unknown outside academia and its potential has not been fully exhausted.

The GIS&T domain is constantly developing further due to scientific and technological advances. An overview of GIScience developments as contributed by Blaschke and Strobl [3] highlights among other topics the potentials of larger data availability in comparison to earlier days of GIScience. Camara et al. [4] discuss the elements of a GIS of the 21st century in comparison to the GIS of the 20th century. They stress the increased importance of sensor networks, mobile devices and remote sensing on the technology side as well as semantics, time and cognition on the concepts side. Their observations include the demand for training GI engineers, who are focused on GI technology development and can collaborate with GI scientists [4]. Their work shows that shaping a domain requires reacting to new developments and adapting educational programs to the requirements of the domain respectively the market.

The BoK cannot be static as technology and science evolve. Several initiatives are working on an update of content and format of the BoK [5-8]. A major joint effort in this direction is currently made under the framework of the European Project "Geographic Information: Need to Know" (GI-N2K). GI-N2K contributes a European perspective to the development of a demand driven GIS&T BoK.

The basis for re-designing the BoK in the GI-N2K project is an assessment of current and future workforce demand and educational supply in the GI domain. This article presents the preliminary results of a survey focusing on workforce demand and aims towards an analysis of the match between the knowledge areas of the current BoK and today's geospatial workforce demands as well as presumed future market trends. Workforce demands are thereby differentiated for different types of organizations and highlight the diversity in levels of expertise in different knowledge areas required by employees.

2 Knowledge Areas of the GIS&T Body of Knowledge

The BoK divides geographic information science and technology into ten Knowledge Areas (KAs) [2]. Each KA covers a set of units that are further subdivided into topics. For each topic the BoK lists learning objectives that are taking four knowledge types into consideration: factual, conceptual, procedural, and meta-cognitive knowledge. The types of knowledge can be related to different levels of cognitive processes such as remember, apply, evaluate, etc., which allows the adaptation of learning objectives for educational programs on different education levels as for Europe defined in the European Qualifications Framework[1]. The level of detail of topics covered by the BoK is extensive. The table below provides only an overview of KAs (first hierarchical level) with some examples of according units (second level) (Table 1). A full version of the BoK can be downloaded from the web[2].

Table 1: Knowledge Areas of the GIS&T BoK (after [2]).

Knowledge Area	Example units included
Analytical Methods	geometric measures, analysis of surfaces, spatial statistics
Conceptual Foundations	philosophical foundations, domains of geographic information, relationships
Cartography and Visualization	data considerations, graphic representation techniques, map production
Design Aspects	project definition, database design, application design
Data Modeling	database management systems, vector and object data models, tessellation data models
Data Manipulation	representation transformation, generalization and aggregation, transaction management
Geocomputation	computational aspects and neurocomputing, cellular automata, heuristics, genetic algorithms
Geospatial Data	map projections, satellite and shipboard remote sensing, land surveying and GPS
GIS&T and Society	legal aspects, dissemination of geospatial information, geospatial information as property
Organizational & Institutional Aspects	origins of GIS&T, managing the GI system operations and infrastructures, coordinating organizations

3 Workforce Demand Assessment

3.1 Aims and Approach

Updating the Body of Knowledge requires a detailed insight in current requirements of the GI job market and foreseeable future developments. An online survey was run by the GI-N2K project in order to assess GIS&T workforce demand. The target group of the survey was people actively working in the GIS&T domain. These people were asked to rate the importance of BoK KAs within their professional life. The intended outcome was job profiles that show required competences and skills of GIS&T in public, private, academic and non-governmental organizations.

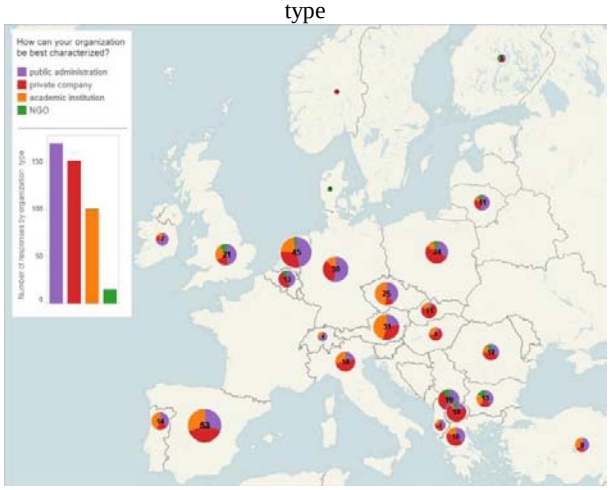
As survey participants were introduced to the BoK by listing KAs, units and exemplary topics (3rd hierarchical level), previous knowledge of the BoK was not required. In terms of content the survey strictly followed the existing KAs in order to avoid predetermining potential adaptations to the BoK. However, within the online survey the KAs itself were presented in random order to ensure approximately equal attention to each KA given the overall length of the survey.

Participants were also asked to name their current job tasks, presumed tasks in five years and individual learning objectives. The collective description of currently performed GIS&T tasks aimed at giving a broad overview of today's workforce, whereas the judgment of future directions was expected to provide opinions on trends in the field. Finally, the educational aims should help to link the workforce demand to an eventual reshaping of educational offers.

3.2 Facts and Figures about the Survey

The online survey was distributed through 31 project partners and networks such as the Association of Geographic Information Laboratories for Europe (AGILE). In total more than 1000 questionnaires were returned out of which 435 were completely filled. Contributions came from over 33 mostly European countries and people working in different types of organizations (Figure 1).

Figure 1: Number of Responses by Country and Organization type



Also the highest level of education in the GI domain was specified by respondents. Following the European Qualifications Framework (EQF) seven levels of expertise

[1] http://ec.europa.eu/eqf/home_en.htm
[2] http://www.aag.org/galleries/publications-files/GIST_Body_of_Knowledge.pdf

were differentiated: beginner, user, competent user (self-trained), competent users (extensively trained), Bachelor, Master and Doctorate. One third of respondents hold a Master's degree in GIS&T. About 12% each are either competent users (self-trained or extensively trained) or have a PhD in GIS&T. The remaining participants hold a Bachelor's degree in GIS&T, are beginners or plain users. The gathered information on organizational affiliation, job description and the educational level of respondents allows a differentiated view regarding the rated importance of KAs.

4 First Survey Results

The presentation of first results focuses on the ratings of the KAs regarding organization type and education level of respondents. The following figures present the mean rating of each KA by category.

Figure 2 shows the mean rating of KAs per type of organization. The mean ratings are similar over organization types for most KAs. The rating given by respondents working in academic institutions differs most from the other categories (the discussed ratings from the academic field are marked with a filled circle). This becomes apparent when on the one hand looking at *analytical methods* and *geocomputation*, which are rated higher by people from the academic field. On the other hand, the two KAs of *GIS&T and society* and *organizational and institutional aspects*, are rated lowest by respondents from the academic field. A detailed interpretation of these results and an assessment of statistical significance yet have to follow.

Figure 3 presents the mean ratings of KAs by people with different levels of educational training in the GIS&T field. The results indicate that the importance of KAs increases with the level of education of the respondents. That means that respondents with a doctorate consistently rate KAs higher than respondents with a Bachelor degree or even lower levels of (mostly informal) GI-education. We attribute this fact to the larger knowledge and experience of highly qualified professionals regarding the topics covered in each KA. This result seems correlated with the rating of KAs through people working at academic institutions.

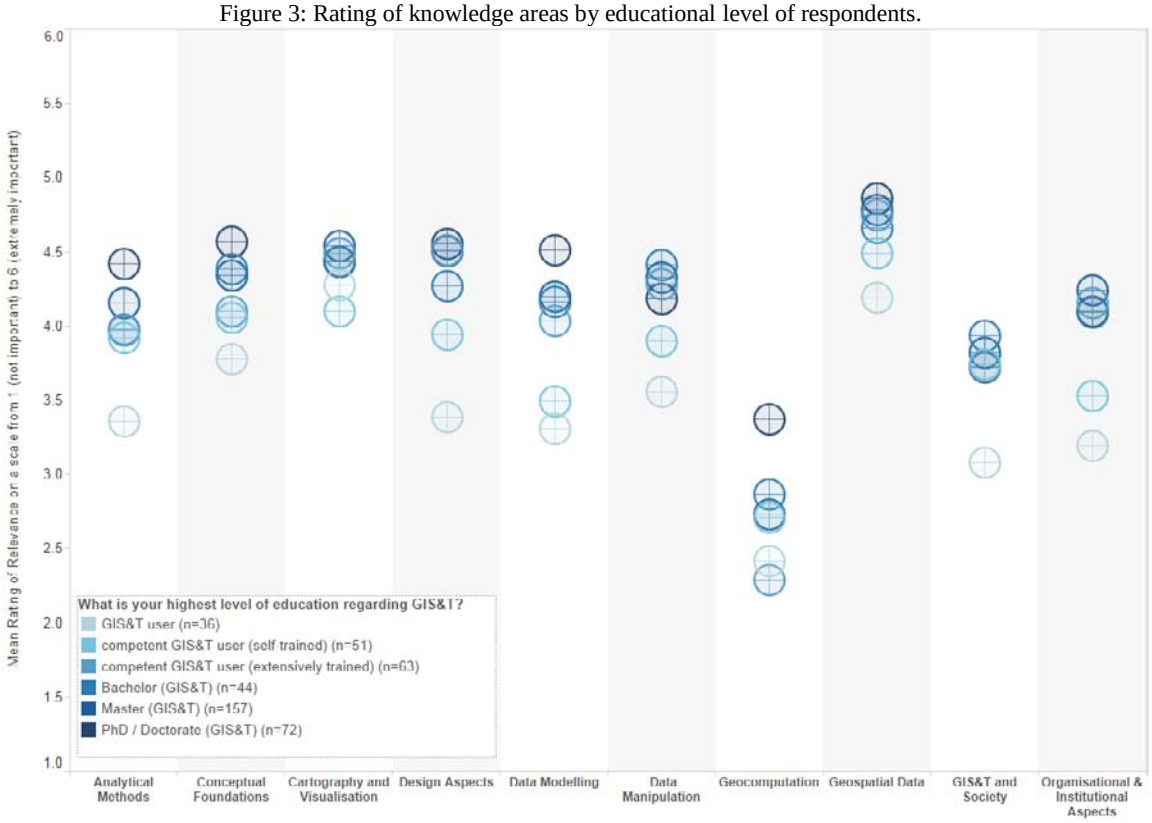
Some KAs are not rated highest by people with a doctorate, but by Bachelor or Master degree holders. An example is the KA data manipulation. However, statistical testing showed that this difference in the rating is not significant.

Comparing the overall ratings of KAs, three KAs are rated considerably less important: *geocomputation*, *GIS&T and society*, and *organizational & institutional aspects*. In the KA *geocomputation*, the concepts and methods covered relate to heuristics, uncertainty, fuzzy sets, cellular automata, agent-based modeling, neurocomputing and others. It can be hypothesized that the sometimes quite advanced concepts covered by this KA are too specialized for tasks in a non-academic yet professional working context.

The other observation is the rating of the KAs *GIS&T and society* and *organizational & institutional aspects*. The ratings differ more across types of organizations and again the overall ratings are lower in comparison to the other KAs. This might be an indication that GIS&T is still primarily seen as a technical discipline.

Figure 2: Rating of knowledge areas by organization type (NGOs have been omitted due to unstable means because of the small sample size).

Figure 3: Rating of knowledge areas by educational level of respondents.



The following figure (Figure 4) shows a word cloud created from free-text replies on the following question: *Which competences will gain importance in the next 5 years?* The keywords GIS and data were removed before generating the word cloud as they dominate the result otherwise. The replies largely point into the same direction that was indicated by [4]. Mobile and web technologies gain importance as well as related topics like applications and development. Analysis, which could mean spatial analysis or data analysis, is expected to continue as important part of GI expertise.

Figure 4: Future competences identified by respondents (the keywords *GIS* and *data* have been excluded from the word cloud).

5 Conclusions and Further Work

The results presented here are preliminary and only indicate the direction in which the detailed analysis of the survey could lead. Clearly and not surprisingly, the competences used in the professional life vary for professionals with different levels of education in the GIS&T domain. Education on different levels clearly steers the profiles of graduates (cf. [9]).

The analysis yet has to dwell deeper into the ratings of knowledge areas and associated units. The detailing of the analysis will include testing for statistical significance for observations made.

The quantitative survey is currently complemented by qualitative interviews with highly reputed representatives of academic, private and public administration organizations across Europe. These qualitative interviews provide insights into required competences on the GI market on a general level.

Next steps are the extended analysis of survey results and the consolidation of findings. The extended analysis will also look into the free text answers given on tasks, trends and learning objectives.

Acknowledgements

This work has been funded by the GI-N2K Project under the European Lifelong Learning Program. Many thanks to the survey participants!

References

[1] D. DiBiase, M. DeMers, A. Johnson, K. Kemp, A. T. Luck, B. Plewe, and E. Wentz, "Geographic Information Science & Technology Body of Knowledge," ed.

Washington, D.C: Association of American Geographers, 2006.

[2] D. DiBiase, M. DeMers, A. Johnson, K. Kemp, A. T. Luck, B. Plewe, and E. Wentz, "Introducing the First Edition of Geographic Information Science and Technology Body of Knowledge," Cartography and Geographic Information Science, vol. 34, pp. 113-120, 2007.

[3] T. Blaschke and J. Strobl. (2010) Geographic Information Science Developments. GIS.Science. 9-15.

[4] G. Câmara, L. Vinhas, C. Davis, F. Fonseca, and T. Carneiro, "Geographical Information Engineering in the 21st Century," in Research Trends in Geographic Information Science, G. Navratil, Ed., ed: Springer Berlin Heidelberg, 2009, pp. 203-218.

[5] M. DeMers, A. Klimaszewski-Patterson, R. Richman, S. Ahearn, B. Plewe, and A. Skupin, "Toward an Immersive 3D Virtual BoK Exploratorium: A Proof of Concept," Transactions in GIS, vol. 17, pp. 335-352, 2013.

[6] I. Hossain and W. Reinhardt, "Curriculum Design Based on the UCGIS S&T Body of Knowledge Supported by a Software Tool," in 8th European GIS Education Seminar, EUGISES 2012, Katholieke Universiteit Leuven, Belgium, 2012.

[7] M. Painho and P. Curvelo, "Building dynamic, ontology-based alternative paths for GIS&T curricula," in Teaching Geographic Information Science and Technology in Higher Education, D. Unwin, N. Tate, K. Foote, and D. DiBiase, Eds., ed: Wiley-Blackwell, 2011, pp. 97-116.

[8] F. I. Rip and E. Verbree, "EduMapping the evolution of an academic GI curriculum – the case of Geomatics at Delft University " in 8th European GIS Education Seminar, EUGISES 2012, Katholieke Universiteit Leuven, Belgium, 2012.

[9] J. Strobl, "Digital Earth Brainware. A Framework for Education and Qualification Requirements," in Geoinformatics paves the Highway to Digital Earth, J. Schiewe and U. Michel, Eds., ed Osnabrück: University of Osnabrück, 2008, pp. 134-138.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Some strategic national initiatives for the Swedish education in the geodata field

Lars Harrie
Karin Larsson
David Tenenbaum
Lund University
Department of Physical Geography and Ecosystem Science
Sweden
karin.larsson@nateko.lu.se
lars.harrie@nateko.lu.se
david.tenenbaum@nateko.lu.se

Milan Horemuz
Royal Institute of Technology
Division of Geodesy and Geoinformatics
Stockholm, Sweden
horemuz@kth.se

Hanna Ridefelt
Gunnar Lysell
National mapping and land registration authority
Gävle, Sweden
Hanna.Ridefelt@lm.se
Gunnar.Lysell@lm.se

S. Anders Brandt
Eva A.U. Sahlin
University of Gävle
Department of Industrial Development, IT and Land Management
sab@hig.se
evasan@hig.se

Göran Adelsköld
Mats Högström
Jakob Lagerstedt
Swedish University of Agricultural Sciences
Uppsala, Sweden
goran.adelskold@slu.se
mats.hogstrom@slu.se
jakob.lagerstedt@slu.se

Abstract

This paper describes national cooperation in Sweden launched by its universities and authorities, aimed at improving geodata education. These initiatives have been focused upon providing common access to geodata, the production of teaching materials in Swedish and organizing annual meetings for teachers. We argue that this type of cooperation is vital to providing high quality education for a poorly recognized subject in a country with a relatively small population.

Keywords: GIS, geodata, education, teaching material, spatial data infrastructure

1 Introduction

This paper describes some strategic national initiatives that have advanced Swedish geodata education during the last decade. The aim of these initiatives has been to improve cooperation between universities that provide geodata education, as well as between universities, authorities and the society. We argue that these cooperative initiatives are necessary to sustain the continuing development of a poorly recognized and financed subject in a country with a relatively small population.

Most universities in Sweden provide short introductory GIS courses. In 2006 the total number of courses were 145 [3], and it is estimated to be similar today. Four universities have study programmes where GIS is a core subject: Lund University (Master's programmes, both campus and distance learning), Karlstad University (Bachelor's programme), the Royal Institute of Technology (Master's programme), and University of Gävle (Bachelor's and Master's programmes).

Besides the specifically GIS-oriented programmes, closely related programmes offered in Sweden are programmes in land surveying, spatial planning, geography and landscape architecture. In land surveying education Lund University and the Royal Institute of Technology have engineering programmes (5 years), and University of Gävle and University West have bachelor of science/engineering programmes.

In addition to the traditional university programmes, there are several shorter GIS and surveying programmes of one or two years in length with the aim of developing professional and practical skills.

Overall, the number of courses in the geodata field has certainly increased during recent decades. But this increase mainly consists of an extensive number of basic GIS courses in programmes where it is not a core subject, and in shorter practical programmes. In education at the master's level, where geodesy, photogrammetry and cartography are the core subjects, availability in Sweden has decreased.

Some trends during the last 15 years that have triggered the national initiatives described in this article are:

- a rapid increase of the utilization of geodata in society
- an increased need in society for competence within GIS, geodata and related topics
- the Bologna process induced student mobility
- the EU directive INSPIRE.

2 National cooperations

2.1 An Educational section in the Swedish Cartographic Society

In 2006 an educational section within the Swedish Cartographic Society was formed. The aim was to fulfil the need for a common platform for educational matters at all levels, and for increased student recruitment within the areas of interest for the Swedish Cartographic Society, such as land surveying, spatial planning, geography and geomatics [4].

An important issue encountered during the first years of the educational section was the operationalization of the Bologna model in 2007. The purpose of this initiative was to make higher education in the EU member states comparable in terms of levels, credits, grades, etc., in order to facilitate mobility of students between universities. In this respect, GIS, being a broad and young subject, is apparently "problematic". As shown in [3], many universities in Sweden at that time provided both GIS-related courses and complete study programmes at both bachelor's and master's level. If and when students wanted to switch programmes or continue their studies at another university, the validation of syllabi in relation to prerequisites for further studies was often difficult. Therefore, in an attempt to improve both cooperation between universities and to produce GIS course syllabi in line with the new Bologna rules, the educational section of the Cartographic Society decided to develop a harmonized syllabus template (see [5]). This template can be used for any subject area where GIS is given as an introductory course. Thanks to the specified learning outcomes, it is now much easier for the receiving university to decide if a student can be admitted to their more advanced courses.

The educational section of the Swedish Cartographic Society also arranges an annual conference for teachers at all levels as well as other interested parties. Here educational matters pertaining to land surveying, spatial planning, geography and geomatics are discussed. The aim of the conference is to facilitate the sharing of experiences, to improve collaboration opportunities, networking and competence development, and to enhance course quality and recruitment of new students. The conferences usually have a special theme each year; recent examples include "the job market and education", "open geodata for education and research", "remote sensing in bachelor and engineering programmes", etc. The conference attracts mainly teachers with GIS interests, which is also reflected in the conference programme.

In order to reach new participants, the location and hosting educational institution of the venue has changed each year. Since 2012, the conference is arranged to coincide with Swedish Map Days every second year, in order to highlight the importance of education at the largest national event within the field. The presentations are normally published on the website of the Swedish Cartographic Society and a short summary of the educational conference is published in the Society's journal Kart & Bildteknik [2, 6, 10].

2.2 Common access to geodata

The production of geodata in Sweden is, as of February 2014, only partly financed via taxes and hence the responsible authorities partly rely on fees to defray costs. With the increased utilization of geodata it was quickly realised that easy access to data for education and research purposes was needed at low cost. Between 2004 and 2011, universities had access to basic geodata for a highly reduced fee through an agreement between the National Library of Sweden and Lantmäteriet (the Swedish mapping, cadastral and land registration authority). Because of organisational and legal changes, this agreement ceased to be in effect in 2012.

In response to the changed situation and ongoing developments regarding data access, both at the European and national levels (EU directive INSPIRE), the Association of Swedish Higher Education and the Swedish Research Council jointly sought a national solution that would ensure that all universities would have continued, easy access to various types of spatial data from national data producers.

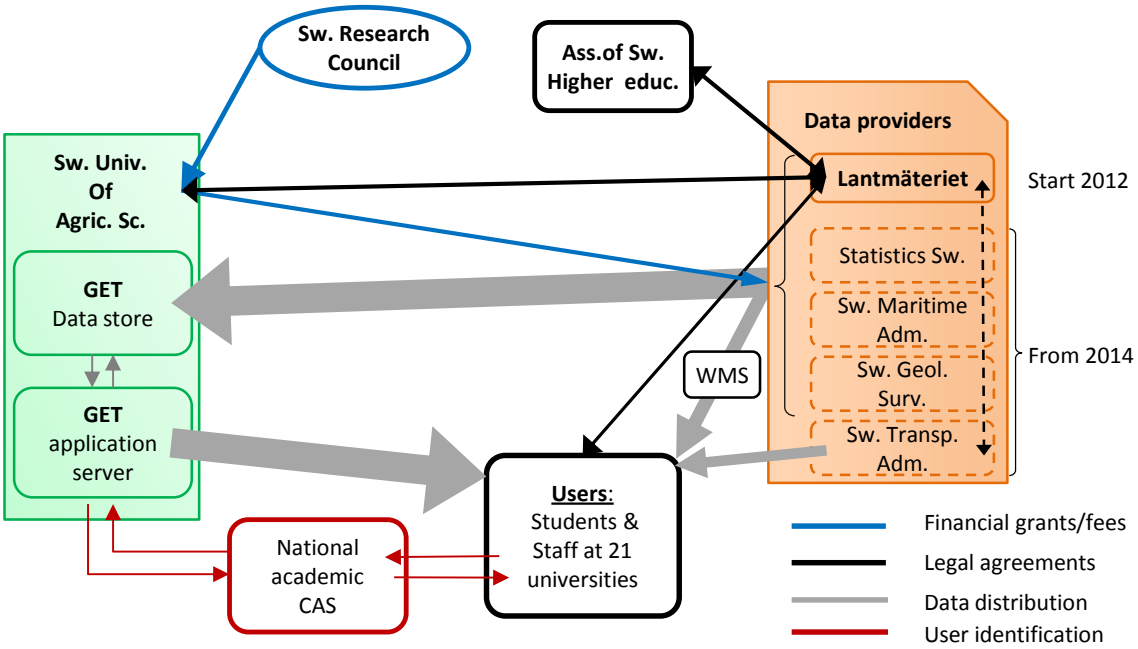
A growing awareness of the importance of geodata resulted in the Swedish Research Council, as part of a national investment in research infrastructure, granting financing for the years 2012-2016. The grant covers data licence fees, distribution service development and the assessment of long term solutions. The basis for implementation is cooperation and the development of common national solutions. It was agreed that the Swedish University of Agricultural Sciences would develop and operate a distribution system that would serve all universities. A prerequisite was the utilisation of existing national infrastructure for user authentication to which almost all universities are connected.

During the initial two years, data from Lantmäteriet, e.g. digital general maps, elevation data, and aerial photographs were included. The distribution system, GET (Geographic Extraction Tool), was successfully developed and made available in 2012 and the service has found widespread use. During the first 16 months of operation the service had about 5,000 unique users from 21 different universities, and the number of downloads were about 1,000 per week. Data and distribution are, among other subjects, discussed in academic users meetings organized by Lantmäteriet and a hosting university twice per year.

During the period 2014-2016 the aims are to (i) enhance and expand the GET service, (ii) incorporate more licensed data, and (iii) evaluate different long-term administrative and financial solutions.

Some examples of geodata that will be incorporated and made available are quaternary deposit mapping from the Geological Survey of Sweden, hydrographical data from the Swedish Maritime Administration, gridded population data from Statistics Sweden, and the National Road Database from the Swedish Transport Administration. In the future, the plan is to further develop the GET application so that service providers can choose between open source and proprietary software when deploying the service.

Figure 1: Technical and organizational outline of data provision via GET and WMS



Looking ahead, the INSPIRE directive will require Government authorities to publish download services for many of their geodatabases. If the GET service is modified to incorporate these download services as data sources, end users would be able to benefit from a familiar interface and access to even more information.

2.3 Facilitating GIS training in schools and innovations based on geodata

Lantmäteriet is leading a project together with the municipality of Västerås with the aim of integrating the use of geodata and spatial analysis in high school education. Since 2010, the national curriculum of geography states that pupils should develop skills related to GIS, and this is therefore the focus of the project's first phase. There is a general lack of knowledge of GIS among upper secondary school teachers, to help support them, the project will provide a website containing both data from national authorities and local municipalities, as well as tutorials and exercises directly related to learning outcomes in the curriculum. A prototype for the site will be available in spring 2014 and will then be tested by teachers.

Thirteen of the national spatial data producing authorities have worked together to plan a public hackathon event, "Hack for Sweden", which will be held in March 2014. The target groups are students, developers and data journalists, who could create new smart services based on spatial data (www.hackforsweden.se).

3 Teaching material in Swedish

Two examples of Swedish literature in the geodata field are described below. The main audience of this literature has been university students but it has also attracted a professional audience, and. is e.g. currently recommended reading Lantmäteriet's "Handbok i mät- och kartfrågor", which is a series of guidelines for surveying and mapping aiming to facilitate standardized handling of geographical data.

3.1 Text book in GIS

In 1995 the first GIS book in Swedish was published [9]. This book, together with English literature, was used in GIS courses in the late 1990s. However, there was a need of a more comprehensive book in Swedish. In 1999 Byggforskningsrådet (a research foundation), ULI (the association for geographic information in Sweden) and the GIS Centre at Lund University published the first edition of the book *Geografisk Informationsbehandling* [7]. The book had nine authors from three universities and the national mapping agency.

The book has since the first edition went through several major and minor revisions. The sixth edition was published in 2013; now by a commercial publisher (Studentlitteratur) [8]. The number of authors had increased to nineteen representing three universities, the national mapping agency, two private companies and one municipality.

There are advantages and disadvantages to having the large number of authors that collaborate to produce this book. The main advantage is that many perspectives are included in the book. This is especially important as the book is used in

courses with diverse student groups (engineers, geographers, forestry and agriculture students, etc.). The main disadvantage is the effort required to reduce heterogeneities in the book concerning content, structure and language, which has made the editing process significantly more difficult.

An important question is whether it is economically viable to publish a comprehensive GIS book in Swedish when there are English alternatives. To make economic sense, the book must be used by all, or at least most, universities that provide basic GIS courses. In this respect, the book has been successful. It has been used by most universities and university colleges in Sweden for more than a decade. The book has sold around 1000 copies each year. This has not been particularly profitable, but it has given the authors and their employers fair compensation for the significant time and effort that has been devoted to the book project.

We believe that there are two strong justifications for continuing to produce a national GIS book. The first is to create and maintain a national, in this case Swedish, vocabulary in the field. Of course, such a vocabulary is also promoted by the Swedish Standards Institute, but our experience is that it is important that a proper vocabulary is provided in basic university education. This is facilitated by the current situation at Swedish universities, where bachelor's education is principally given in Swedish, and subsequent master's education in English.

The second reason to have a national GIS book is that some of the pertinent content is truly national in nature. For example, the book has one chapter describing Swedish geographic data infrastructure and another contains description of the national geodetic reference systems. Also, a substantial part of the practical examples in the book are nationally-sourced; this familiarity of context has pedagogic value for Swedish students.

3.2 National compendium in geodesy and photogrammetry

There are several types and levels of education in Sweden, such as vocational training, bachelor's and master's programmes, where surveying courses are part of the curriculum. In many cases, older or obsolete literature has been used, which does not reflect the latest trends and technical innovations in surveying. Moreover, the terminology and level of detail are different in different programmes. This situation provided the main motivation to write a new compendium [1], covering the most important topics in surveying. There are eight authors; four of them are active lecturers at Swedish universities (Royal Institute of Technology, University of Gävle and Lund University), two authors work at Lantmäteriet and two are working at private companies (Blom and Tyréns). The production of the compendium was financed by the employers of the authors and by a contribution from the Swedish Cartographic Society.

The compendium is written in Swedish and covers the following topics:

- geodetic reference systems
- cartographic projections
- terrestrial surveying instruments and methods
- uncertainty in measurements and the least-squares method
- GNSS surveying
- photogrammetry
- terrestrial and airborne laser scanning.

The compendium is published electronically and protected by a Creative Commons by-nc-nd license, which means that anybody can copy and redistribute the compendium for non-commercial purposes, but it must be properly referenced and is not allowed to be modified. There are two main reasons why it was decided to publish the compendium electronically: to maximize availability and flexibility. As the main aim was to produce a common textbook in basic surveying suitable for all relevant programmes in Sweden, it was a natural choice to publish it on Lantmäteriet's web page. Electronic publication also has the advantage of easy updating. All source files are available to all authors on-line via a cloud service. Currently the Royal Institute of Technology is responsible for update management.

We have strived to spread the compendium as widely as possible. Therefore, we have chosen to use a Creative Commons license, rather than a book with copyright. The market for literature in geodesy and photogrammetry is also considerably smaller than for GIS, and it would probably not have been economically feasible to spend the extra time to create a book based on the compendium.

4 Concluding remarks

Education in the geodata field has undergone some significant changes in Sweden. There are more universities providing courses in the field (mainly basic GIS courses) at the same time as education offerings on master's level are on the decline. Meanwhile there is an increased need in society for competent personnel. To cope with this situation, cooperation between the universities and with other national bodies, e.g. geodata producers, is vital. In this paper we have described some common initiatives, including teaching material, development of course syllabi, common portals and financial solutions for geodata access, as well as teachers networking. We believe that this cooperation is essential to maintaining the quality of geodata education in Sweden. One indication that geodata education has achieved a good level of quality was the quality assurance evaluation in 2012-2013 by the Swedish Higher Education Authority. In this evaluation, all science and engineering educational programmes were evaluated, and the results were overall good for the programmes related to geodata.

References

[1] B. Andersson, A Boberg, L Harrie, M. Horemuz, P.-O. Olsson, C.-G. Persson, Y. Reshetyuk, and H. Rost. *Geodetisk och fotogrammetrisk mätnings- och beräkningsteknik*. Nnational Swedish compendium in S urveying. KTH, Lantmäteriet, Högskolan i Gävle, Kartografiska sällskapet and Lund University, 2010.

[2] J. Bohlin. Rapport från årets utbildningskonferens (in Swedish). *Kart och Bildteknik (Mapping and Image Science)*, 2011(3):18, 2011.

[3] S. A. Brandt, J. M. Karlsson and P. Ollert-Hallqvist. Harmonization of GI educations in Sweden and the Bologna process - viewpoints of University of Gävle. In *Proceedings of the Fifth European GIS Education Seminar (EUGISES 2006)*. September 7-10, 2006, Cracow-Pieniny Poland, 2006.

[4] S. A. Brandt and A. Larsson. Kartografiska Sällskapets utbildningssektion – ett nytt tillskott i KS-familjen (in Swedish). *Kart & Bildteknik (Mapping and Image Science)*, 21(2):10-11, 2006.

[5] S. A. Brandt and W. Arnberg. A harmonized GIS course curriculum for Swedish universities. In *EUC'07 HERODOT Proceedings, ESRI European User Conference*, Stockholm, Sweden, 25-27 September 2007.

[6] S. A. Brandt. Rapport från Utbildningssektionens årliga utbildningskonferens (in Swedish). *Kart & Bildteknik (Mapping and Image Science)*, 2009(3):18-19, 2009.

[7] L. Eklundh, editor. *Geografisk informationsbehandling – metoder och tillämpningar* (in Swedish), 1tt edition. Byggforskningsrådet, Stockholm, 1999.

[8] L. Harrie, editor. *Geografisk informationsbehandling – teori, metoder och tillämpningar* (in Swedish), 6th edition. Studentlitteratur, Lund, 2013.

[9] B. Malmström, and A. Wellving. *Introduktion till GIS*, ULI, Gävle, 1995.

[10] R. Nyberg. Rapport från årets lärarkonferens (in Swedish). *Kart & Bildteknik (Mapping and Image Science)*, 2010(3):18, 2010.

Session:
Linked Data Web

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Cadastral data integration through Linked Data

Jhonny Saavedra
Universidad Politécnica de Madrid
Madrid, Spain
ja.saavedra@alumnos.upm.es

Luis M. Vilches-Blázquez
Ontology Engineering Group, Dpto. de Inteligencia Artificial, Facultad de Informática.
Universidad Politécnica de Madrid, Madrid, Spain
lmvilches@fi.upm.es

Alberto Boada
Instituto Geográfico Agustín Codazzi
Bogotá, Colombia
aboadar@igac.gov.co

Abstract

Cadastral data is one of the more important types of geospatial data. Taking into account the importance of these data, several international bodies have worked for creating a standardised model for land administration. However, in spite of existing efforts, there are several open issues for the development of a harmonized vision of cadastral data. Taking into account this scenario, Linked Open Data may allow addressing some of these challenges, by proposing best practices for exposing, sharing, and integrating data on the Web.

This paper shows a use case where two cadastral information sources are semantically integrated according to Linked Data principles. These sources belong to different Colombian cadastral producers and are characterized by different heterogeneity issues. Herein, we describe an implementation of Linked Data principles in the cadastral domain using LADM standard (ISO 19152) and GeoSPARQL. Besides, our original data are enriched with different dataset of Linked Data cloud (LinkedGeoData and GeoNames).

Keywords: Cadastre, Linked Data, Land Administration Data Model, GeoSPARQL.

1 Introduction

Cadastral data is one of the more important types of geospatial data. They have been the basis for the land tributes since Roma's times. These data are defined as the geographic extent of the past, current, and future rights and interests in real property including the spatial information necessary to describe that geographic extent [1]. Rights and interests are the benefits or enjoyment in real property that can be conveyed, transferred, or otherwise allocated to another for economic remuneration. Rights and interests are recorded in land record documents. The spatial information necessary to describe rights and interests includes surveys and legal description frameworks such as the Public Land Survey System, as well as parcel-by-parcel surveys and descriptions [1].

Taking into account the importance of these data, several international bodies have worked for creating a standardised model for land administration. Thus, the Federation International of Surveyors (FIG) started to work at 1996 for developing a future cadastral system (Cadastre 2014) [2]. This proposal, which has become the inspiration of the modern cadastres, has allowed increasing associated services of cadastral systems through technology deployment. Likewise, other more recent efforts for developing standardized models in order to facilitate cadastre data exchange are INSPIRE Data Specification on Cadastral Parcels [3] and the ISO standard 19152 Land Administration Domain Model (LADM) [4].

However, in spite of existing efforts, there are several open issues for the development of a harmonized vision of cadastral data. Some of these challenges are related to integration process of different cadastral data and how to connect these data with related information (e.g.: public services, demographic statistics, planning, etc.).

Linked Open Data may allow addressing some of these challenges, by proposing best practices for exposing, sharing, and integrating data on the Web [5]. The principles of Linked Data were first outlined by Berners-Lee in [6], using the following four guidelines: (1) Use URIs as names for things. (2) Use HTTP URIs so that people can look up those names. (3) When someone looks up a URI, provide useful information, using standards, such as: Resource Description Framework (RDF) and SPARQL Query Language for RDF (SPARQL) and (4) Include links to other URIs, so that they can discover more things. Further details about sets of rules for publishing data on the Web are shown in (Berners-Lee 2006).

Several approaches generating and publishing geospatial Linked Data are appearing in the state-of-the-art in order to perform a semantic information integration process. The capabilities of LOD for integration and interoperability into geospatial context have been recognised by many authors like [7], [8], and [9]. This, and fact attaches an increasing the interest for publishing geospatial information as Linked Data over the last years. Nowadays, we find more than 68 geospatial datasets and six billion of triples with location (often, a pair of coordinates) in the Linked Data cloud.

This paper shows a use case where different cadastral information sources are semantically integrated according to Linked Data principles. These sources belong to different Colombian cadastral producers and are characterized by different heterogeneity issues. Herein, we describe one of the first implementations of Linked Data in the cadastral domain using LADM standard (ISO 19152) and GeoSPARQL[1]. Besides, our original data are enriched with different dataset of Linked Data cloud (LinkedGeoData and GeoNames).

[1] http://www.opengeospatial.org/standards/geosparql

This paper is structured as follows. We start providing a description of our use case for integrating Colombian cadastral data (section 2). Next, we present a brief overview of the existing related work (section 3). In section 4, we describe the process for generating and publishing cadastral Linked Data. Finally, we summarize some conclusions and identify future work in section 5.

2 Colombian Cadastre: An integration use case

Colombia's cadastre has been one of the most developed in Latin-American Region. Currently, the National Cadastral Authority is the Colombian National Geographic Institute (*Instituto Geographic Agustin Codazzi* – IGAC). Furthermore, this country has four additional cadastral producers, which generate information in a decentralised way, in the municipalities of Medellin, Bogotá, Cali and Antioquia.

These different cadastral producers have different and heterogeneous models, vocabularies, and their own production and management systems. A representative example of this heterogeneity is associated with the National Authority, which has different cadastral model in order to manage generated data.

In order to overcome these barriers, IGAC is working on a Cadastral National System. The goals of this project are to create and consolidate a unique cadastral model for National data, deploy this model in a distributed database system and to create a web for providing these data to final users. However, this project has not taken into account the decentralised cadastral offices, due to the fact that there was not an agreement between several producers for centralizing the cadastral data management.

Driven by this scenario, we present a use case, where main purpose is to support the reusing, exchange and semantic integration of Colombian's cadastral data. Within this use case, which focuses on physical aspect of cadastral information, we deal with heterogeneous datasets, which belong to different producers, and they are reused and semantically integrated in order to connect cadastral data and keep provenance of the different sources and producers. This diversity of producers and datasets entails different issues related to heterogeneity of datasets, which are solved using Linked Data principles. Thus, we take cadastral datasets from the National producer (IGAC) and another dataset from a local producer, concretely from Bogotá cadastre[2].

3 Related work

Currently, there are more than 890 RDF datasets tagged as Linked Data in the Datahub[3]. From these datasets, 61 are tagged as geographic data[4]. These data not only come from geospatial research labs, otherwise they are published by important producers, such as Ordnance Survey[5], National Geographical Institute of Spain[6], U.S. Geological Survey[7], and so on. Besides, there exist important collaborative initiatives (e.g.: OpenStreetMap[8] and GeoNames[9]) that are part of this movement. Within published geospatial Linked Data the most common topics are related to transport, toponyms, administrative boundaries, environmental and statistical data. There are no data associated with cadastral information in this repository (Datahub). However, there exist an initiative related to register information called "The Land Register of UK[10], which is publishing the register and cadastral information according to Linked Data. Besides, in [11] is described an approach for cadastre-register integration using ontologies and Semantic Web.

With respect to cadastral models, Land Administration Domain Model (LADM) - ISO 19152 is the most recent resource in the cadastral topic. This standard provides terminology for land administration and enables the combining of land administration information from different sources in a coherent manner. In spite of the fact that it is a recent standard, LADM has already profiles in different countries, such as: Spain, Portugal, Germany, and Japan among others. Taking into account LADM standard, there is a first approach for developing an OWL ontology of LADM [10].

4 Cadastral data integration

In order to perform this integration process, we use the methodological guidelines for generating, integrating and publishing geospatial data according to Linked Data principles described in [12]. These guidelines propose an iterative incremental life cycle model where data gets continuously improved and extended. It consists of the following steps: (1) specification, (2) modelling, (3) RDF generation, (4) links generation, (5) publication, and (6) exploitation. The detail of each of them is shown in the next items.

4.1 Specification

As aforementioned, we work with datasets from a National producer (IGAC) and a local producer (Bogotá cadaster). Next, we describe main characteristic associated with these datasets: On the one hand, the cadastral data of IGAC are stored in an ESRI geodatabase. Within this geodatabase, we work with a subset of cadastral data, which was provided by IGAC in shapefiles (*shp*). These shapefiles belong to the municipality of *Soacha*. On the other hand, Bogotá cadaster data are available in its website in different formats (e.g.: SHP, KMZ, DWG or GML). In the context of our work, we downloaded and manipulated data in shapefile format. In this

[2] The used dataset in this work belongs to Bogotá Spatial Data Infrastructures (IDECA). http://www.ideca.gov.co/

[3] The Datahub is a data management platform from the Open Knowledge Foundation which collects many of the data sets published as Open Linked Data. http://datahub.io/es/dataset?tags=lod

[4] http://datahub.io/es/owns/dataset?tags=geographic&tags=lod

[5] data.ordnancesurvey.co.uk/

[6] http://geo.linkeddata.es/

[7] http://cegis.usgs.gov/ontology.html

[8] http://linkedgeodata.org/

[9] http://www.geonames.org/ontology/documentation.html

[10] http://landregistry.data.gov.uk/

case, we work with a subset of two localities of the Bogotá city (a subset related to the central area of Bogotá and another, which limits with *Soacha* municipality, called *Bosa*). Both of these datasets have attributes related to their information domain (e.g.: area, name, label, geometry, address, etc.). However, they also have differences in their models and used vocabulary.

With respect to URI design, we adopt recommendation of [13], that is, we design our URIs to be simple, stable and manageable. Taking into account this, we use the domain http://datos.igac.gov.co/ for publishing our cadastral Linked Data. According to this root domain, we propose the following URI pattern for this work: In order to identify the provenance of data, we create two URI for pointing to different producers. Thus, IGAC resources are identified with *http://datos.igac.gov.co/id/catastro/igacsnc/*+ and Bogotá resources use the following URI *http://datos.igac.gov.co/id/catastro/bogota/*+.

Finally, with respect to the URI pattern for resources (instances), we use the following pattern, where we add to each URI the National cadastral code:

- IGAC:
  http://datos.igac.gov.co/id/catastro/igacsncl/
  25754010100000069000500000000
- Bogotá cadastre:
  http://datos.igac.gov.co/id/catastro/bogota/
  25754010100000045290200010000

4.2 Ontology modelling

For the modelling of the information contained in the aforementioned datasets we have created an ontology network, which is a collection of ontologies joined together through a variety of different relationships such as mapping, modularization, version, and dependency relationships. This network has been developed following the NeOn methodology [14], by reusing existing ontological and non-ontological resources. Next we provide some details about used resources.

Regarding cadastral domain, we reuse a non-ontological resource, which is the core of our ontology network. This resource is the ISO 19152 Land Administration Domain Model (LADM). LADM proposal has not a deep level of detail and is composed of two classes (*SpatialUnitGroup* and *SpatialUnit*).

Taking into account the general viewpoint provided by LADM, we decided to perform an extension of this proposal for considering Colombian cadastral characteristics. Thus, we developed a profile of LADM, called LADM_CO, for modelling these issues in an ontology. A subset of the LADM_CO ontology is shown in the Figure 1, which includes classes as *neighbourhood*, *block*, *construction*, *land*, and so on. Besides, we developed two different ontologies for modelling characteristics of each Colombian cadastral producer (IGAC and Bogotá cadastre). After we developed these ontologies, our work focused on setting mapping between the components of these ontologies.

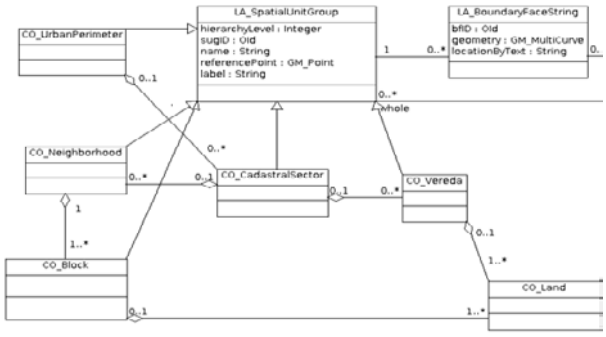


Figure 1. Main classes of the LADM_CO ontology

.

Likewise, we developed an ontology for modelling Colombian administrative boundaries, and reuse GeoSPARQL ontology in order to achieve two essential issues for our work. On the one hand, being able to represent complex geometry and, on the other hand, supporting spatial queries for exploiting cadastral information.

4.3 RDF generation

According to the followed methodology, in this activity we have to take the data sources selected in the specification activity, and transform them to RDF according to the vocabulary mentioned in the modelling activity. We use different systems to transform cadastral features from aforementioned datasets into RDF.

For dealing with geospatial information, we develop an extension of Geometry2RDF[11], called shp2GeoSPARQL[12], in order to transform geometrical information from datasets to RDF. This extension parses shapefiles in order to retrieve the associated geometric data, and generate the geospatial RDF according to GeoSPARQL ontology. For dealing with thematic data, we use Google Refine[13] and its RDF extension[14].

[11] http://oeg-upm.net/index.php/en/technologies/151-geometry2rdf
[12] https://github.com/jasaavedra/shp2geosparql/
[13] https://code.google.com/p/google-refine/
[14] http://refine.deri.ie/

Listing 1. Example RDF of building IGAC

```
<http://datos.igac.gov.co/id/catastro/igacsnc/257540000000000110039000000000> a snc:RConstruccion ;
        snc:codigoConstruccion "257540000000000110039000000000" ;
        snc:tipoConstruccion "CONVENCIONAL" ;
        snc:tipoDominio "PRIVADO" ;
        snc:numeroPisos "2"^^xsd:int ;
        snc:numeroSotanos "0"^^xsd:int ;
        snc:numeroMezanines "0"^^xsd:int ;
        snc:numeroSemisotanos "0"^^xsd:int ;
        snc:identificador "A" ;
        snc:codigoEdificacion "1"^^xsd:int ;
        snc:codigoAnteriorConstruccion "25754000000110039000" ;
        snc:area "48.1300000037695"^^xsd:double ;
        snc:perteneceA <http://datos.igac.gov.co/id/catastro/igacsnc/257540000000000110039000000000> .
        snc:hasgeometry <http://datos.igac.gov.co/id/catastro/igacsnc/Geometry/237f9dc77f73be36d9371c2420c01e37cac362ac>

<rdf:Description rdf:about="http://datos.igac.gov.co/id/catastro/igacsnc/Geometry/237f9dc77f73be36d9371c2420c01e37cac362ac">
    <rdf:type rdf:resource="http://www.opengis.net/ont/sf#MultiPolygon"/>
    <geo:asWKT rdf:datatype="http://www.opengis.net/ont/sf#wktLiteral">MULTIPOLYGON (((984419.3488769521 998229.4562988281,
</rdf:Description>
```

This RDF generation process allows transforming data of dealt datasets into structured and non-proprietary data. Besides, RDF data are generated according to a common and shared vocabulary (developed ontology network) and, therefore, they are semantically interoperable. An example of the generated RDF is shown in the Listing 1. This RDF is associated with a land parcel of Bogotá. In this RDF we include information about points of interest from LinkedGeoData and geometrical information serialized in WKT. Besides, we include topological relationships in the aforementioned RDF, due to the fact that we add information associated with block and adjoining parcels.

4.4 Links generation

We cadastral data was linked with two useful datasets, such as: LinkedGeoData and GeoNames.

Taking into account that LinkedGeoData and GeoNames data are represented by points and these datasets have no cadastral attributes, location information is the only one that we can use in the linking process. This fact does not allow using traditional linking tools, for instance, SILK framework[15]. Therefore, we perform this process using a GIS tool and spatial analysis to determine that points of GeoNames and LinkedGeodata were contained in each land parcel of the Colombian cadastral data. This analysis was performed previously RDF generation process. Thus, we generate a file with our URIs and use Google Refine and its RDF extension for setting links with the aforementioned datasets. This means that the RDF and linking generation was performed at the same time in the context of our work.

4.5 Publication

We have published our datasets using Parliament[16], which was chosen as triple store because it deals with GeoSPARQL standard and can be used to make spatial query in order to exploit our data.

When the data was loaded in our triple store, we performed some tests for checking the consistency of transformation process and quality of results. By this way, we solved minor mistakes related to RDF data and our ontology network. Some examples of the executed queries executed are shown in the Listing 2, and the Figure 2. In Listing 2 we check the consistency of the transformation process in each dataset using the developed ontology network. In this example we ask about number of buildings with more than of 5 floors within a specific area. This query allows us checking that we integrate both datasets in a right way.

Listing 2: SPARQL query for checking transformation process

```
SELECT DISTINCT
count(?id)
WHERE {
   ?id rdf:type   snc:UConstruccion.
   ?id ladm:namespace "IGAC".
   ?id snc:numeroPisos ?NoPisos.
   filter (?NoPisos > 5) .
} 73
```

```
SELECT DISTINCT
count(?id)
WHERE {
   ?id rdf:type   bog:Construccion
   ?id ladm:namespace "UAECD".
   ?id bog:numeroPisos ?NoPisos.
   filter (?NoPisos > 5) .
} 1414
```

```
SELECT DISTINCT
count (?id)
WHERE {
   ?id rdf:type ladmco:Building.
   ?id ladmco:floorNumber ?NoPisos.
   filter (?NoPisos > 5)
} 1487
```

Furthermore, due to the fact that linking process performed, we enrich our original data. This enrichment of our data is exploited through an SPARQL query (Figure 3). With this query we add to our data addresses associated with different points of interest, which are gathered from GeoNames and LinkedGeodata.

Fig 2: SPARQL query for checking links

```
SELECT
?nombre ?direccion ?nbarrio

WHERE {
   ?s bog:puntoDeInteres ?PI .
   ?s rdf:type <http://linkedgeodata.org/ontology/Shop> .
   ?s rdfs:label ?nombre .
   ?s bog:tiene ?tiene .
   ?tiene bog:direccion ?direccion.
}
 limit 20
```

nombre	direccion
"Crayon"	"CL 18A 1 11"
"Seneka"	"CL 18A 1 83"
"Terraza Pasteur"	"CL 24 6A 19"

15 http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

16 http://parliament.semwebcentral.org/

5 Conclusions and future work

In this paper we described the process followed to generate, integrate and publish cadastral Linked Data from two Colombian producers. The main goal of this work was to allow combining different sources using Linked Data principles, for overcoming current problems of information integration associated with National producers of this information type.

For achieving our goal, we have developed an ontology network, which is reusing LADM standard and GeoSPARQL ontology, and have generated an extension of Geometry2RDF tool for dealing with characteristics of our datasets. Besides, we have integrated and enriched Colombian cadastral data with two different datasets of the Linked Data cloud (GeoNames and LinkedGeodata). It demonstrates that interaction with other kinds of data is possible too.

Future work will continue integrating more datasets from other cadastral producers through Linked Data principles. We will also focus on identifying and interlinking other cadastral features with knowledge bases belonging to the Linked Open Data Initiative. Furthermore, we will work on exploitation process in order to show our cadastral Linked Data in a friendly way for final users. Finally, we will focus on improving existing metrics for linking process in order to deal with characteristics of cadastral information and increase the accuracy of this process and, therefore, existing tools.

Acknowledgements

We would like to kindly thank all members involved in this work, especially people from IGAC and Bogotá Spatial Data Infrastructures (IDECA) for their interest, help and support.

6 References

[1] Federal Geographic Data Committee. (2008). Geographic Information Framework Data Content Standard – Part 1: Cadastral. https://www.fgdc.gov/standards/projects/FGDC-standards-projects/framework-data-standard/GI_FrameworkDataStandard_Part1_Cadastral.pdf

[2] Federation International of Surveyors. (1998). Cadastre 2014: A vision for future cadastral systems. http://www.fig.net/cadastre2014/translation/c2014-english.pdf

[3] INSPIRE. (2009). Data Specification on Cadastral Parcels – Guidelines. http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_CP_v3.0.pdf

[4] ISO. (2012) ISO 19152, Geographic information — Land Administration Domain Model (LADM). http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51206

[5] Heath, Tom and Bizer, Christian. (2011). Linked Data: Evolving the Web into a Global Data Space. http://linkeddatabook.com/editions/1.0/

[6] Berners-Lee, T. (2006) Linked data. World Wide Web design issues. http://www.w3.org/DesignIssues/LinkedData.html.

[7] Sheth, A. P., (1998) Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In Interoperating Geographic Information Systems, Goodchild, M. F., Egenhofer, M. J., Fegeas, R., Kottman, C. A. (eds), pp. 5–30, Kluwer.

[8] Goodchild, M., Egenhofer, M. J., Fegeas, R. Kottman, C. Eds. (1999) Interoperating Geographic Information Systems. The International Series in Engineering and Computer Science, Kluwer.

[9] Kuhn, W. (2005) Geospatial Semantics: Why, of What, and How? In Spaccapietra, S. Zimányi, E. (Eds.): Journal on Data Semantics III. Lecture Notes in Computer Science, 3534 (3), 1-24.

[10] Kean Huat Soon (2013) International FIG workshop on the Land Administration Domain Model. http://wiki.tudelft.nl/pub/Research/ISO19152/ImplementationMaterial/LADMOntology.owl

[11] Piña, Nelcy et. al., (2011). Ontología web semántica del registro catastral venezolano. http://cesimo.ing.ula.ve/~jacinto/websemantica/Art%C3%ADculo%20Ontolog%C3%ADa%20web%20sem%C3%A1ntica%20del%20Registro%20Catastral%20Venezolano.pdf

[12] Vilches-Blázquez, L.M., Villazón-Terrazas, B., Corcho, O., Gómez-Pérez, A.: Integrating geographical information in the Linked Digital Earth. International Journal of Digital Earth. ISSN 1753-8947. (2013).

[13] Sauermann et al. (2006). Cool URIs for the Semantic Web. http://www.dfki.unikl.de/~sauermann/2007/01/semweburisdraft/uricrisis.pdf

[14] Suarez, Maria, et al, (2010). NeOn Methodology for Building Ontology Networks. http://oa.upm.es/3879

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

GEOSUD SDI: accessing Earth Observation data collections with semantic-based services

Mathieu Kazmierski
UMR ESPACE-DEV (IRD)
mathieu.kazmierski@ird.fr

Jean-Christophe Desconnets
UMR ESPACE-DEV (IRD)
jean-christophe.desconnets@ird.fr

Bertrand Guerrero
UMR ESPACE-DEV (IRD)
bertrand.guerrero@ird.fr

Dominique Briand
UMR ESPACE-DEV (IRD)
dominique.briand@ird.fr

Abstract

Ecosystems and territories are complex systems involving multidisciplinary approaches on different time scales and different locations. Yet, mastering the spatial information on these systems is critical to lead a relevant environmental research, as well as addressing efficient public policies. Although the volume of Earth Observation (EO) data generated these last years greatly increased, their usages are still too limited when it comes to environmental issues. So as to remedy to this underutilisation, the GEOSUD (GEOinformation for Sustainable Development) project had undertaken to deploy a national Spatial Data Infrastructure (SDI) to ease access to high resolution and very high resolution satellite images for public stakeholders and scientists. This paper gives an overview of this infrastructure and places emphasis on the innovative components that fit the specific needs in terms of images discovery of the GEOSUD community. These components relies on domain controlled vocabularies, which can assist both experts and non-experts in the field of remote sensing in their search of the appropriate material to suit their needs.

Keywords: satellite images, discovery services, metadata, faceted search, geoprocessing

1 Introduction and background

Ecosystems and territories are complex systems requiring multidisciplinary approaches on different time scales and different areas. Yet, mastering the spatial information on these systems is critical to lead a relevant environmental research, as well as addressing efficient public policies. Considering this statement, European initiatives such as the INSPIRE directive or the even broader Open Data initiative[1] have been set up in the past few years and are now well established. Thus, spatial data usage of vector data from large and well-known repositories has considerably developed within the targeted community of users in the past few years.

Earth Observation (EO) data strengthen these repositories and are offering observations data at relevant spatial and spectral resolutions with high acquisition frequency that eventually allow to carry specific studies on dynamics of territories. Although the volume of EO data generated these last years greatly increased [1], their usages by public stakeholders and scientists are still limited when it comes to environmental issues.

The core problems are well identified. The first concerns the high costs for user licence of satellite images. Indeed, many offers for high resolution or very high resolution images require to pay for using images with a quite restrictive and expensive licence that eventually brings quite substantial financial costs.

Besides, the lack of awareness of what is on offer, regarding to the amount of satellite images, and the varying degrees of capacity and knowledge skills in the field of remote sensing of end-users make their choice of a sensor and associated product challenging. In fact, there are numerous dedicated applications for discovery and access to satellite images although they are provided with hardly comprehensible user interfaces for non-expert audiences. Finally, the multiplicity and the lack of standardisation in nomenclatures (e.g. the multiple names of a processing level depending of the image provider) as well as in image descriptions are major obstacles for users to easily access to a clear view of the wide range of imagery products available on a given territory and to quickly evaluate if it fits their needs.

Issues related to the access of distributed and heterogeneous data are quite common. Usually, it is solved by the deployment of a Spatial Data Infrastructure [2,3]. The COPERNICUS[2] initiative on a European level and the GEOSS [4] on a global level had implemented this principles and give now access to products on a regional, continental or global scale.

In France, the GEOSUD project started in the finding that public stakeholders working in the field of environmental management and public policies underuse satellite images. It had undertaken to deploy equivalent measures as the ones stated above to ease the access to high resolution and very high resolution EO data for public stakeholders and scientists. This project began in 2011 and is led by a consortium of 12 organisations among which public structures, universities, research institutes, companies and spatial data end-user communities. In addition to the acquisition of national annual high resolution coverages the first five years, the satellite images offer will be broadened by a receiving antenna GEOSUD that will allow to program and acquire images from different types of high resolution or very high resolution sensors.

The main goals of the GEOSUD project are to guarantee and ease the access to satellite images, by simplifying their use licences, the discovery and the download of its resources, and in a second phase, by guaranteeing access to on-line geoprocessing that serves the working domains of GEOSUD end-users through an image analysis application. So as to fit the needs of a heterogeneous users community, regarding

[1] European Open Data: https://open-data.europa.eu/en/data/

[2] COPERNICUS : http://www.copernicus.eu/

their comprehension level of remote sensing concepts, this project has to ensure that the different user interfaces and services are adapted to the various degree of expertise.

Moreover, the GEOSUD SDI will be one part of the institutional sector dedicated to satellite image access (called the "Pôle THEIA"). The latter aims at providing a wide range of satellite data on continental surfaces [5]. These data are produced thanks to different projects funded within the French scientific community, among which the main ones are GEOSUD, Postel, Kalideos, Hydroweb, Take Five, Spirit,…

The THEIA infrastructure will be built as a federation of data and services centres for satellite images, in respect of each access conditions and for a broader targeted audience than GEOSUD. Common services are in the heart of this federation, including in particular an image discovery application. It aims to give a unique access point to all available data in the federation, with increased transparency. An identification and authentication common mechanism is considered. The GEOSUD user database would have to be interoperable with the latter.

This paper presents the GEOSUD SDI. In particular, we focus on the innovative components that meet the specific needs in terms of data access and data discovery for non-expert end-users. These semantic components rely on a set of controlled vocabularies.

The paper is organised as follows. After a brief presentation of the GEOSUD context and remind the fundamental principles underlying this SDI in the Section 2, the Section 3 details the two main innovative components of the SDI: the data standardisation and semantic annotation service, and the data discovery application, which make use of annotations to facilitate both the discovery process and the image selection process for end-users. The Section 4 gives an overview of the technical choices that will be implemented in the GEOSUD infrastructure this year. The section 5 concludes this paper by reminding all the expected benefits from this infrastructure for GEOSUD end-user community, the contribution of this SDI in the national infrastructure THEIA as well as of the expected use of high performance computing for large-scale geoprocessing that will be handled as the next step toward innovative services for public policies.

2 Interoperability of access services

The principles that led to the design of the GEOSUD infrastructure is based on the definition of a spatial data infrastructure as proposed by the INSPIRE directive: « the metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use ...operated or made available in an interoperable manner »[6]. Thus, a SDI that follows these rules must give access to data through discovery, visualisation and download interoperable services.

The adoption of international standards, that both ensure data harmonisation and access services standardisation, allows on one hand the aggregation of heterogeneous data sources from multiple satellite images providers and, on the other hand, the unification of their description so as to offer a broad and homogeneous vision of available data.

The ISO Technical Comity TC/211 and the Open Geospatial Consortium (OGC) are the main designers of the standards in the field of spatial data and services. In the particular context of Earth Observation data, spatial agencies such as ESA (European Spatial Agency) have highly contributed to define these specifications (e.g. Heterogeneous Mission Accessibility specifications).

To provide images both to the European environmental management community and to the EO community, we have committed ourselves to take into account the recommendations of the INSPIRE directive as well as the specifications emitted by the EO community when we designed the access components and the underlying metadata models. As for the visualisation and download services, they were designed according the OGC standards: WMS (Web Map Service), WMTS (Web Map Tile Service) for the first one; WCS (Web Coverage Service) for the second one.

3 Enhancing images discovery by enriching metadata from heterogeneous sources

Images discovery web-services make use of the information contained in the metadata. Most of the existing web-services for images discovery, since they are based on standards, whether from OGC, as the Catalog Service for the Web (CSW) standard [7], or as OpenSearch with its EO extension (EO OpenSearch) [8], use a reduced set of metadata. On the one hand, this reduced set does not reflect the richness offered by metadata of image providers. On the other hand, it often offers unsatisfactory expressiveness to build requests on a specific characteristic of an image e.g. its spatial resolution or spectral bands. It also limits the results filtering and ranking, which are critical factors when the web-service gives access to a large number of images. Moreover, it offers little if any metadata on the image semantic, which may be of key importance for the selection process depending on its intended purpose [9].

3.1 Abstract metadata model for EO metadata insertion

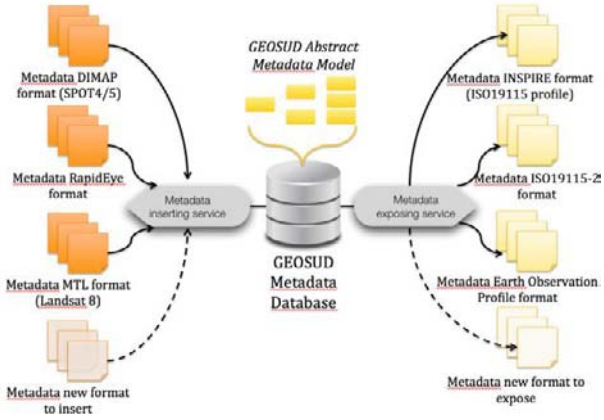
Image providers are given metadata in non-generic models: DIMAP for SPOT and Pleiades images, MTL for Landsat. Other producers adopted metadata description standards such as ISO 191115-2 [10] or the OGC Earth Observation Profile [11].

Moreover, the GEOSUD SDI addresses as well to the EO users community than to other thematic communities. Then it must provide metadata and interoperable discovery services for these communities, by assuring access to metadata compliant with both INSPIRE and OpenSearchGeo Spatial and Temporal Extensions specifications.

In this context, the multiplicity of input and output formats requires many transformations. Many methods are offered to deal with heterogeneous metadata. This reference [12] describes some of them to assure metadata interoperability at the schema level. The switching-across method appears to be the most efficient in our case. It has been developed from the crosswalking method, which consists in the mapping of syntactic and semantic elements from one model to another.

The latter works well when the number of metadata models is relatively low, what is not the case in our context. So as to make the crosswalking more efficient when input and output models are numerous, the switching-across method consists in channelling transformations through a switching schema from the input models to the output models. By doing so, it limits the amount of transformations by avoiding model-to-model mapping. Thus, so as to minimize costs and effort of transforming metadata, the adopted method is to base the transformations on an abstract model, which is given a "switching-across" role (see Figure 1). In later stage of the project, the insertion of images from new sensors will be possible and will be eased by this approach.

Figure 1: Abstract metadata model GEOSUD for the insertion and export of metadata in various standardised or not models



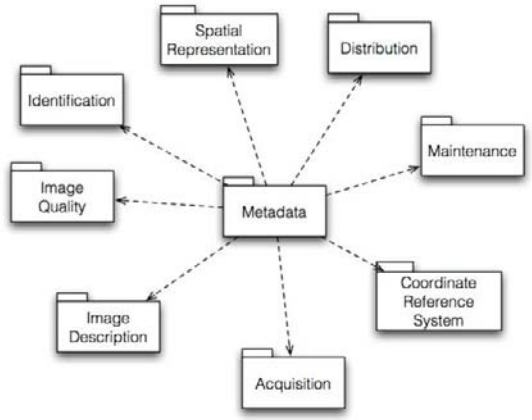
Source: UMR ESPACE-DEV, IRD

In the Figure 2, we present a general view of the GEOSUD abstract model. So as to cover a broad range of functionalities (discovery, visualisation, processing, perennial archiving), it gives a great deal of information. It covers information on image identification such as its *geographicalExtent*, its description (*imagingCondition*, *processingLevelCode*, *pixelResolution*), its acquisition condition (*GSD_instrument*, *GSD_SpectralBand* classes). It also provides content on the evaluation of the image quality (*GSD_QuantitativeResults*, *GSD_QualityReportDocument* classes) and their lineage (*GSD_Lineage* class).

It allows designing lasting components to insert metadata in the SDI and will ease the metadata insertion from new sensors.

The abstract metadata model is deployed through the metadata insertion service. Mapping schemes between source models (DIMAP, MTL) to the abstract model are also provided. The insertion service read the producers metadata and execute the mapping for each metadata according to its native model. The resulting metadata is stored in a database and exposed on demand in an interoperable format (ISO19115 INSPIRE, EOP) through standardised web-services that fits the user needs.

Figure 2: GEOSUD abstract model (packages view)



Source: UMR ESPACE-DEV, IRD

So as to illustrate the use of the abstract model in its switching schema role, we give an extract of a mapping between models, implemented in the insertion service of the GEOSUD SDI (see Table 1). These mappings assure the transformation of elements based on the DIMAP model used to describe SPOT or PLEIADES images toward the ISO 19915 INSPIRE model, so to as to expose metadata through a CSW 2.0.2 AP ISO discovery service.

3.2 Automatic annotation of images metadata

The image metadata provided by data producers deal essentially with their intrinsic features such as their footprint, represented as a polygon or a bounding box, their acquisition date or the spectral bands that compose image. They also deal

Table 1: Extract of crosswalks between DIMAP model (SPOT, PLEIADES products), GEOSUD abstract Model and ISO 19115 model

DIMAP elements	Geosud Abstract Model elements	ISO 19115 elements
Min(Dataset_Frame/Vertex/FRAME_LON) Min(Dataset_Frame/Vertex/FRAME_LAT) Max(Dataset_Frame/Vertex/FRAME_LON) Max(Dataset_Frame/Vertex/FRAME_LAT)	GSD_Identification.geographicalExtent	MD_DataIdentification.extent.geographicElement
Dataset_Sources/Scene_Source/INSTRUMENT Dataset_Sources/Scene_Source/INSTRUMENT_INDEX	GSD_Intrument.instrumentShortName	N/A
Production/PRODUCT_INFO	GSD_ImageDescription.processingLevelCode	MD_ImageDescription.processingLevelCode
Dataset_Sources/Scene_Source/MISSION_INDEX Dataset_Sources/Scene_Source /SENSOR_CODE*	GSD_GridSpatialRepresentation.pixelResolution	MD_Resolution.distance

Source: UMR ESPACE-DEV, IRD

with the image acquisition and image production conditions, such as the processing level (e.g. 1A, 2A, 2B...).

Consequently, if an expert in the field of remote sensing can achieve to search efficiently these images by relying on its knowledge, most end-users will not, since their knowledge of concepts or specific vocabulary would be insufficient.

Adapting a "provider" vocabulary to a "consumer" vocabulary may be necessary to enhance users search experience.

It is indeed simpler to make a request based on a toponym like "I am looking for all the images that cover the city of Toulouse" than to draw a bounding box using its coordinates. In the same way it is often more relevant to give the possibility to "look for all the images containing urban area" when the search purpose is to look for urban dynamics assessments.

To ensure the vocabulary adaptation and the enrichment of metadata, we rely on internal and external controlled vocabulary. To adapt footprints, we rely on the GEONAMES ontology [13]. It gives access through a REST service to all toponyms across the French territory. For example, when inserting a SPOT5 image with the bounding box {north: 44.49321, south: 43.81890, east:-0.26412, west: -1.21798} into the SDI, the insertion service execute the following HTTP request to the GEONAMES API :
http://api.geonames.org/search?north=44.49321&south=43.81890&east=-0.26412&west=-1.21798&username=geosud

It returns all the toponyms and extra-information on each one of them within the specified bounding box (see Table 2).

Table 2. Example of a Geonames API response in XML format to an HTTP search request

Line number	XML response extract
1	**<geonames** style="MEDIUM">
2	...
3	**<geoname>**
4	<**toponymName**>Mont-de-Marsan</**toponymName**>
5	<**name**>Mont-de-Marsan</**name**>
6	<**lat**>43.89028</**lat**>
7	<**lng**>-0.50056</**lng**>
8	<**geonameId**>6433897</**geonameId**>
9	<**countryCode**>FR</**countryCode**>
10	<**countryName**>France</**countryName**>
11	<**fcl**>A</**fcl**>
12	<**fcode**>ADM4</**fcode**>
13	</**geoname**>
14	...
15	**</geonames**>

Source : Geonames

The latter response is then consumed by the insertion service. It extracts the content from the <toponymName> tag and populates fields of the image metadata GEOSUD database. In this case, the toponym name "*Mont-de-Marsan"* is inserted into *GSD_Identification.geographicIdentifier* field.

Based on the same principle, the enrichment of metadata with land cover information relies on the Corine Land Cover 2006 classification [14]. Adapting vocabulary on spatial resolution or on processing level will be taken in charge by another component of the SDI, which will execute these operations simultaneously to the metadata indexing.

3.3 User faceted search application

Discovery applications on which is based the image search are usually complex for they often offer an expert approach to the search process, where a large number of search criteria are offered through non-intuitive interfaces. Moreover, the semantic of the criteria is not always readily understandable for end-users. To overcome these limitations, we choose to base the search process on an interactive filtering mechanism to retrieve information, which is widely used on the Internet: the faceted search [15,16].

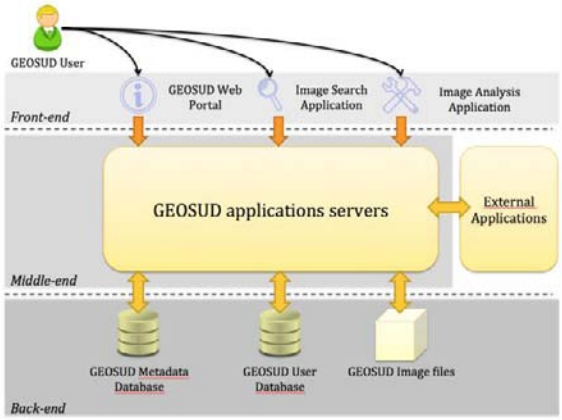
Also referred as faceted navigation or faceted classification, the faceted search is defined as a method to access a data collection by allowing user to explore the latter through selected filters. It is based on a classification system in where multiple categories can be assigned to the same data and where the filtering can be enabled in different ways [17]. In our example, a filter could be an image property : acquisition date, spatial resolution, location,…

The enrichment of metadata and the adaptation of vocabularies during the metadata insertion phase discussed above provide categories for faceted search with a less expert-oriented semantic, close to the various audiences of GEOSUD. For example, we provide a hierarchical facet "spatial location" which rely on administrative toponyms such as region or city names, from which have been extracted the metadata value g*eographicIdentifier*.

4 Implementation

So as to deploy these services under the conditions emitted above and to allow external applications to access in a controlled way to GEOSUD data, the logical architecture of GEOSUD SDI will used a widely accepted 3-tier architecture principles (see Figure 3): front-end user applications, middle-end services that gives access to data and a back-end composed of databases and the image files.

Figure 3: Simplified view of the GEOSUD 3-tier architecture



Source: UMR ESPACE-DEV, IRD

5 Conclusion

High resolution and very high resolution EO data have become essential to undertake environmental research and address efficient public policies. The GEOSUD SDI brings an original and comprehensive solution for public stakeholders and scientists by allowing them to access in a standardised way to satellite images from a wide range of sensors.

One of the original aspects of this infrastructure is that it focuses and adapts to the various degree of expertise of its end-users, which is also a major constraint to search efficiently images in their everyday work. The adaptation and enrichment of metadata from image providers by the use of controlled vocabularies (GEONAMES, Corine Land Cover) allow the search process to share a semantic that is close to a non-expert user. This also helps to build a discovery application that is based on these vocabularies. The choice of a faceted-centred discovery mechanism will enhance the user experience and increase the relevance of returned results.

Today, the user needs consist in the exploitation of images. Thus, their analysis is used to build complex environmental indicators that require specific tools (ENVI, eCognition) as well as large computational and data resources. Yet, these tools are still out of reach of a large part of public stakeholders or scientists. Consequently, the next challenge for GEOSUD SDI is to give access to a satellite image-processing platform that fits the latter audience needs with specific processing chain (e.g. detection of nitrate-fixing intermediate crops). For this purpose, an online computational platform combined with high performance computing environment is considered. In this context, we will attach importance to tackle the barriers created by the various level of expertise of GEOSUD end-users, either in the geo-processes discovery or their configuration and execution.

References

[1] A.J. Tatem, S.J. Goetz; S.I. Hay. Fifty years of earth observation satellites. American Scientist, 96(5):390-398, 2008/9.

[2] Global Spatial Data Infrastructure : Developing Spatial Data Infrastructures: The SDI Cookbook. Version 2.0. Consulted the 25 January 2004 at : http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf

[3] Yang C., Raskin R., Goodchild M., Gahegan M. : Geospatial Cyberinfrastructure : Past , present and future. *Computers, Environment and Urban Systems* 34(2010). 264-277.

[4] Christian E.J. : GEOSS Architecture Principles and the GEOSS ClearingHouse. *IEEE systems journal*, 2(3). September 2008.

[5] Theia Land data center: M.Leroy, P.Kosuth, O.Hagolle, S.Cherchali, P.Maurel, J.Desconnets. ESA Living Planet Symposium. Edimburgh, UK. 9-13 September 2013.

[6] European Parliament, 2007. DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Available at: http://eurlex.europa.eu/JOHtml.do?uri=OJ:L:2007:108:SOM:EN:HTML.

[7] OGC : OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile, Rapport, Open Geospatial Consortium, 2007.

[8] OGC 10-032 : OpenSearch Geo Spatial and Temporal Extensions. 2014

[9] Boisson P., Clerc S., Desconnets JC, Libourel T. : Using a semantic approach for a Cataloguing Service. OTM workshops (2) 2006: 1712-1722. LNCS Springer Heildelberg.

[10] ISO TC/211 : ISO 19115-2 - Geographic information — Metadata —Part 2, Extensions for imagery and gridded data. First edition. 2009

[11] OGC OGC10-157r3 - Earth Observation Metadata profile of Observations & Measurements Standard, 2012.

[12] Chan L.M., Zeng M.L. : Metadata interoperability and standardisation – a study of methodology Part 1. Achieving interoperability at the schema level. *D-Lib magazine* 12(6) ISSN 1082-9873. June 2006

[13] Geonames API web service : http://www.geonames.org

[14] Corine Land Cover WFS web service : http://sd1878-2.sivit.org/geoserver/wfs?

[15] Uddin, M.N., Janecek, P.: Faceted classification in web information architecture: A framework for using semantic web tools. *Electronic Library*, 25(2), 2007

[16] Denton, W.: How to make a faceted classification and put it on the web. See http://www.miskatonic.org/library/facet-web-howto. html (2011).

[17] Laporte, M.-A., Mougenot, I., & Garnier, E. (2013) A faceted search system for facilitating discovery-driven scientific activities: a use case from functional ecology. Workshop S4BioDiv, ESWC 2013, CEUR-WS.org.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Little Steps Towards Big Goals. Using Linked Data to Develop Next Generation Spatial Data Infrastructures (aka SDI 3.0)

Francis Harvey
University of Minnesota
Minneapolis, MN, USA
fharvey@umn.edu

Jim Jones
Westfälische Wilhelms Universität,
Münster, Germany
jim.jones@uni-muenster.de

Simon Scheider
Westfälische Wilhelms Universität Münster, Germany
simon.scheider@uni-muenster.de

Adam Iwaniak
Wrocław University of Environmental and Life Sciences
Wrocław, Poland
adam.iwaniak@up.wroc.pl

IwonaKaczmarek
Wrocław University of Environmental and Life Sciences
Wrocław, Poland
iwona.kaczmarek@up.wroc.pl

JaromarŁukowicz
Wrocław University of Environmental and Life Sciences
Wrocław, Poland
jaromar.lukowicz@struktura.eu

Marek Strzelecki
Wrocław University of Environmental and Life Sciences
Wrocław, Poland
marek.strzelecki@up.wroc.pl

Abstract

Society is moving at an increasing pace toward the next stage of the information society through linked data. Among the relevant developments in geographic information science, linked data approaches offer potential for improving SDI functionality [12]. Linked data uses Semantic Web technologies and makes it possible to link at a very granular level data resources of the web for a multitude of purposes. While the technological implementation in many ways is still in a phase of adolescence, vast amounts of data, including geographic information (GI) have been prepared, for example by the UK Ordinance Survey [8] and other governmental and non-governmental bodies. The overwhelming focus has been on producing RDF formatted data for linked data applications--the foundation for applications. In this short paper, we provide an overview of potentials of linked open data for SDI 3.0 developments. Through two exemplary use cases we illustrate specifically some first steps towards a more web-oriented and distributed approach to creating SDI architectures. The cases demonstrate applications based on the LOD4WFS Adapter, which opens the way for multi-perspective GI applications, created on-demand from multiple GI data resources. These applications automate geometry-based selections of data using spatial queries with the use of RCC8 and OGC Simple Features topological functions. Future work in this area includes adding semantic operators to refine GI processing with multiple ontologies.

1 Introduction

The information age offers a promise of improved access, efficiency, and new discoveries through information. Partly realized in Web 1.0 and 2.0 services and applications, Semantic Web technologies, linked data concepts, fundamental to Web 3.0 implementation, define the next big goals for the information society. Geospatial information occupies an important component of developments involving Linked Data and the Semantic Web. Many initiatives, both governmental and non-governmental continue to develop linked data resources and applications [28]. A number of commercial applications have also embraced these Semantic Web technologies. This article considers these developments points to the potentials of Linked Data for future SDI architectures and broader support.

These developments continue efforts to enhance online information access and use of geographic information, opening the doors to application potentials that only a few decades ago would have seemed to come directly from science fiction. We now have the capability to access GI from anywhere on the globe. Realistically, this potential faces many technological and organizational challenges. Achieving improved access through Semantic Web applications that can handle the semantic issues [9, 15, 16, 10] offers interesting means to support uses hitherto constrained by web 1.0 and 2.0 technologies [2]

Semantic Web technologies, understood as an important part of Web 3.0, in summary, serve as integrators across different content, information applications and systems. Already the applications are diverse with implementations in government [24, 8], commercial applications, entertainment, education [14] and other domains [18]. Linked open data (LOD) in particular refers to the practices for publishing and connecting data on the Web [3]. In other words, linked data offers the technologies for creating dynamic integration on the Web.

If we take a step back, we can see that many Semantic Web technologies offer information integration capabilities envisioned in SDI concepts [22, 20, 19, 31] but challenging to realize. The SDI implementations could only support links at

the file level: each ftp:, http:, https: etc access retrieves an entire file. The operating system-based distribution of files requires additional steps to process data from the Web. A skilled and usually also knowledgeable user needs to locate, download, then prepare all required resources in order to answer a query in the common Web architecture used by SDI. This is clearly complex, time-consuming and inefficient for information integration [23]. The intent of SDI was to support feature level queries and processing, however, querying across repositories distributed on the Web is not feasible. This becomes possible using Semantic Web technologies, such as RDF, and opens possibilities to integrate or aggregate subsets of datasets based on logical criteria, e.g, topological operators determining if features are within a specified extent.

This paper describes a potential bridge between Web 3.0 based-architecture and SDI-type repositories to help make SDI-type applications more efficient and flexible. The proposed connection closely corresponds to OGC goals and standards' development that implement Web 2.0 technologies, e.g., WMS, WFS [30]. LOD additionally provides the foundation for incorporating semantic considerations in GI processing.

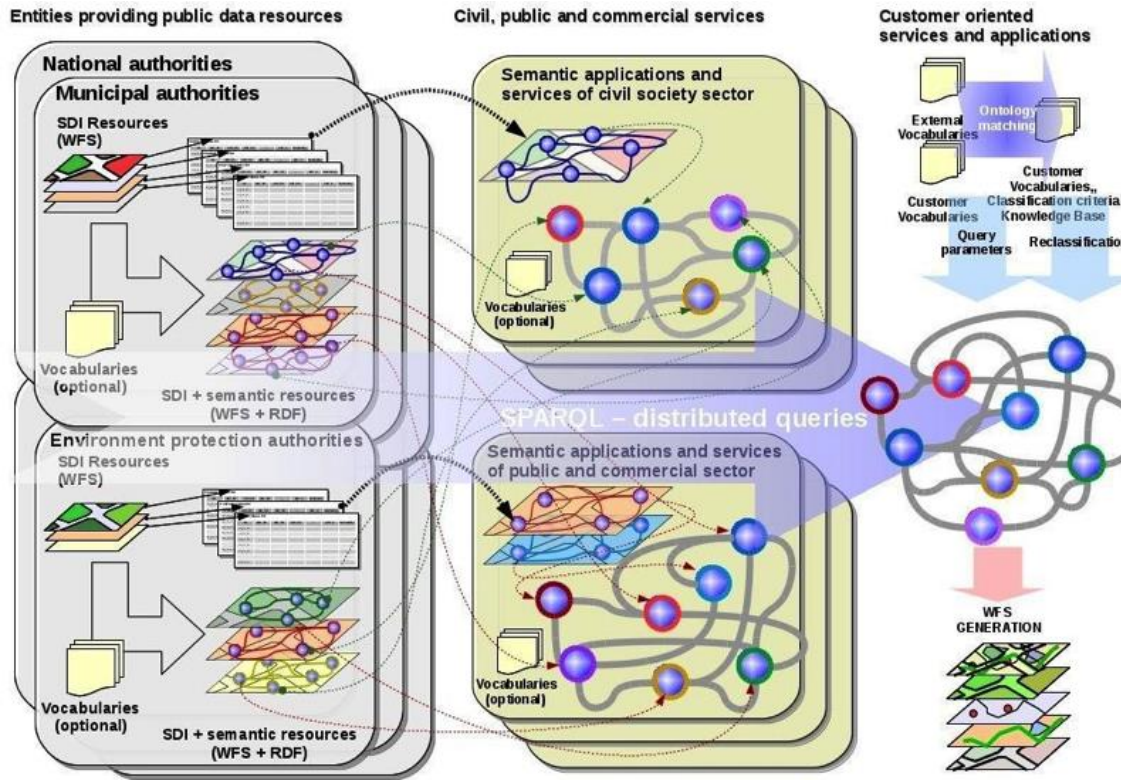
In this paper we consider some LOD related technologies and how we can utilize common technologies to improve feature-level queries over distributed datasets. We focus particularly on developments using the LOD4WFS Adapter, which enables access to LOD data sources via OGC Web Services [13].

2 Technologies, Applications, and Institutions

Geographic applications and information needs range greatly [31] but every GI-Application needs data; and the data is needed by groups. Often portions of single or multiple files are required. Presently, most GI processing requires use of a GIS and/or SDI and proficiency in several areas beyond geoinformatics. The predominant implementation using Web 1.0 and 2.0 approaches only allows for the access of entire files. We need to consider technologies, applications, and institutions holistically to connect uses to data.

The main objective of SDI is to facilitate access to spatial data services through search and preview functionalities provided by portals. Public administration is the main actor as a provider of SDI [11] which, according to the INSPIRE directive, for example, is responsible for the publication of Web services according to OGC standards. The task of public administration is to publish structural data including spatial data, but in a LOD approach this occurs in a way that supports the ability to integrate data from multiple heterogeneous sources using Semantic Web technology, particularly RDF. Consuming RDF data is currently more complex for users compared to the well-known and widely implemented WMS and WFS services. Using the LOD4WFS Adapter we can connect the LOD, SDI repositories, and support multiple GI-Applications (see Figure 1).

Figure 1: Resources, services, and applications illustrating the capability to dynamically construct geographic information graphs to support multiple uses



Retrieving data from SDI resources into RDF triples can be done in two ways. First is a simple transformation. This transformation resembles a W3C Direct Mapping for the transformation from relational model into RDF graph. We use a similar mapping which binds WFS/GML objects to RDF nodes, WFS/GML attributes to RDF graph edges (properties) pointing to literals (OWL datatype properties), references between WFS/GML objects to RDF graph edges pointing to other resources (OWL object properties). The second possibility is associated with the use of ontology languages (e.g. RDF Schema or OWL Tbox) and a mapping ontology. It allows developing more sophisticated transformations, which create a "view" on original SDI resources in terms of an RDF representation. This is compatible to the W3C R2RML adapter (http://www.w3.org/TR/r2rml/). The result of a SPARQL query in this environment is a new graph that aggregates elements from multiple RDF data sources and constructs new connections between them (see Figure 1).

Ontologies occupy a very important role in the nascent web of data. They provide information about properties which relate resources (nodes) in a graph. We can use the NeoGeo vocabulary (NeoGeo Vocabulary, 2014) to supplement repository and data set level ontologies. It makes it possible to formulate SPARQL queries to discover required data from multiple RDF data sources.

A distributed SPARQL query results in an RDF graph consisting of objects retrieved from various sources provided by separated institutions and organizations (see Figures 2 and 4). Using spatial information in the form of semantic representations we can discover spatial relations between heterogeneous data and enrich result graphs with new relations (see Figure 3). New graphs could include new literal values too, which were, e.g., derived from information transformations [4]

One of the main problems connected with using Semantic Web technologies with GI is the heterogeneity of the semantic and spatial data [9, 6, 7, 26, 29, 25]. To properly operate between these two different approaches it is needed to establish bridges. The first need is to be able to serve spatial data as semantic data with the use of RDF data model and appropriate vocabularies or ontologies. There are applications capable of exposing spatial data (geometry and properties) as RDF triples, which can work as online triple store server, e.g. Parliament [1], SemGeo, Strabon [17], and offline converter of spatial data TripleGeo (TripleGeo). Online triple stores also give user the possibility of querying graphs with the use of SPARQL queries with extensions like GeoSPARQL, which provide spatial analysis functions. The second need is to use semantic data within GI systems, which comes down to converting spatial data representation from RDF triples back to GIS-compliant representations like GML or Shapefiles. This part can be done with the LOD4WFS Adapter, which is capable of explore existing triple stores with spatial features and expose them as OGC WFS service, which can be opened and displayed directly in any GIS that implements the OGC WFS standard (Standard Data Access). The LOD4WFS Adapter also provides the possibility of creating on-demand WFS layers from SPARQL queries and executing them on distributed remote triple stores (Federated Data Access). If the query result contains spatial data, it can be converted on the fly into GML and served through a WFS service.

Figure 2: Dereferencing URIs of resources stored in graph for multiple purposes

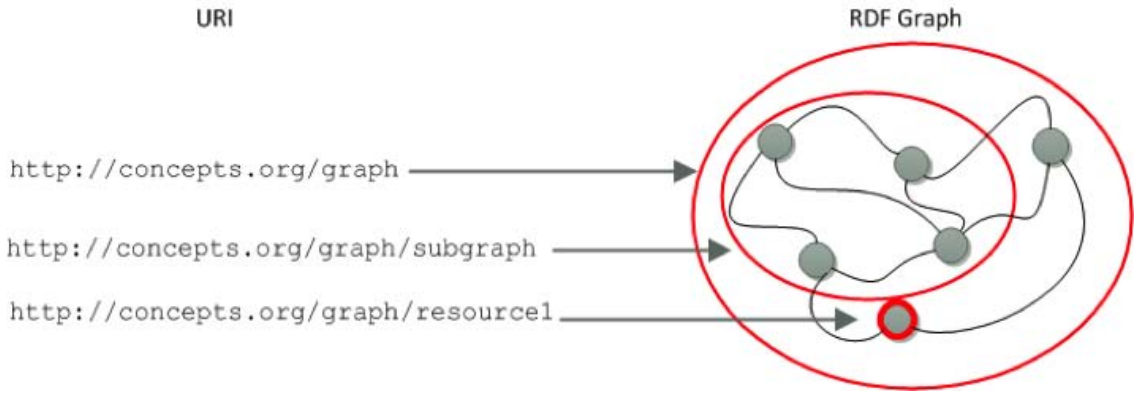


Figure 3: Example of a distributed query integrating data from two different sources.

```
PREFIX : <http://www.example.com/>
SELECT  ?o1 ?o2
WHERE
[
SERVICE <http://example1.com/sparql_endpoint1>
{
  ?s1 ?p ?o1 .
{SERVICE <http://example2.com/sparql endpoint2>
{
    ?s2 ?p ?o2 .

]]
```

The result of the query is the list of East England administrative districts, which is published through a WFS service using LOD4WFS Adapter (Figure 5).

Figure 4: Example of federated query with DBpedia and Ordnance Survey Linked Data.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?abstract ?resource ?name ?entry (concat("POINT(",xsd:string(?long)," ",xsd:string(?lat),")")
AS ?wkt) ?gss ?unitid
WHERE {
SERVICE <http://data.ordnancesurvey.co.uk/datasets/os-linked-data/apis/sparql>
 { ?x <http://www.w3.org/2000/01/rdf-schema#label> ?name.
   ?x <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?lat.
   ?x <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?long.
   ?x <http://data.ordnancesurvey.co.uk/ontology/admingeo/gssCode> ?gss.
   ?x <http://data.ordnancesurvey.co.uk/ontology/admingeo/hasUnitID> ?unitid.
   ?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://data.ordnancesurvey.co.uk/ontology/admingeo/District>
  SERVICE <http://dbpedia.org/sparql/>
   { ?entry <http://www.w3.org/2000/01/rdf-schema#label> ?place.
     ?entry <http://dbpedia.org/ontology/abstract> ?abstract.
     ?entry <http://dbpedia.org/ontology/isPartOf> <http://dbpedia.org/resource/East_of_England>
     FILTER langMatches(lang(?place), "EN")
     FILTER langMatches(lang(?abstract), "EN")
     FILTER ( str(?place) = ?name )
   }}
```

3. Use cases

Semantic web technologies can use linked data for a large range of geographic information application. We consider two use cases in some detail to illustrate the potential of LOD and semantic web technology for extending and improving SDI functionality in both the administrative and civil society domains.

3.1 Supporting administration procedures

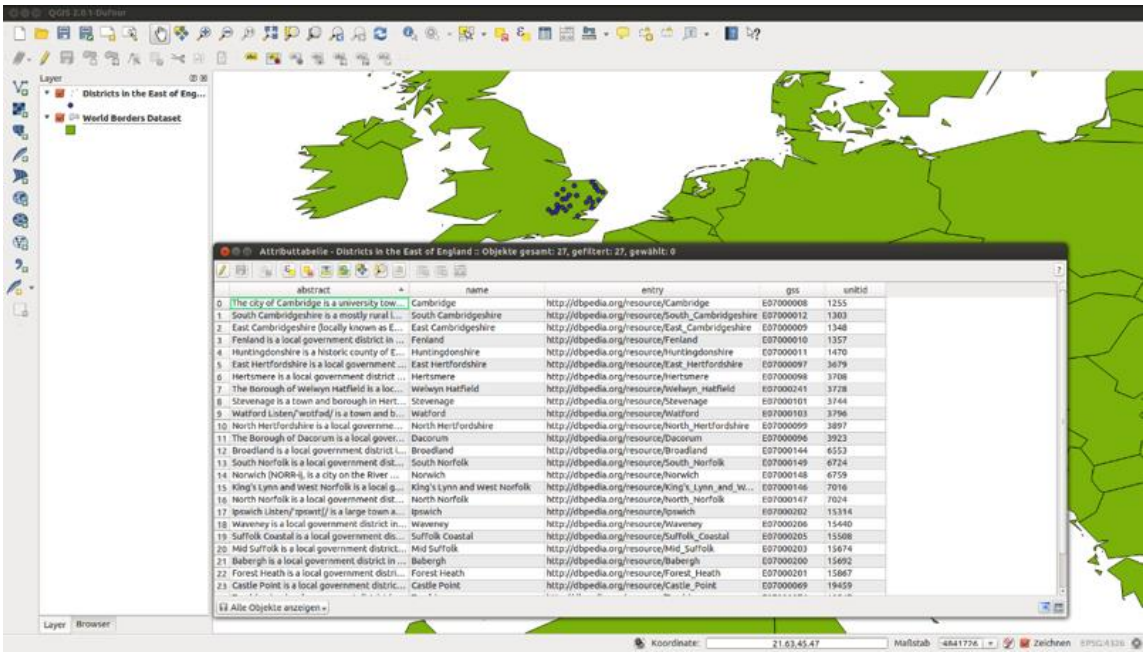
The first SDI-related case to consider is building permit application processing. The case focuses on providing simple access to relevant data by theme and by extent and legal requirements. In the prototype version of the application, administrative staff has to manually determine which data sets to include; the linked data implementation can automatically determine the extent and only access and download the corresponding data from multiple data sources. Later versions of the application can add richer semantic functionality to automate the selection of additional data sets based on criteria and relevant rules and laws.

Most important in the first version of the application is the support of the administrative officer, who must prepare a complete set of required information for review. The decision maker should be able to analyze and check all circumstances relating to the decision making procedure. This person can obtain information to determine,

- if a parcel in question is covered by local development plan,
- if a parcel has appropriate access to local transportation infrastructure,
- if it is possible to access all other required infrastructure systems (water supply, sewage system, electric power supply)
- if the proposed building includes measures to assure public safety

Important in the review and decision process is finding the data sets relevant to the planned uses and long-term

Figure 5: GML data converted by LOD4WFS adapter and published in WFS



developments of a given project. In particular, it is crucial to determine what rules and requirements apply and consider existing easements and liens. When the proposed building is multi-functional, multiple inter-related rules and regulations often apply.

Using a semantic web approach, the data needed for the review and decision come from heterogeneous sources: data repositories, specifications, guidelines, etc. They include public registers, cartographic and cadastral data, information about infrastructure facilities, planning regulations, past building activities and proposals. The application supports review through a browser-based user interface that allows the administrative staff to define geographic location and manually select relevant layers; the decision making process uses the same application architecture enhanced with annotations from the reviewing staff members. The application retrieves RDF resources using the *GeoSPARQL vocabulary*[1] and RCC8 operators [5] through SPARQL queries and assembles them into the linked data resource for supporting the query. Using the LOD4WFS Adapter, linked data can be transformed into GML and distributed through WFS distribution to GIS and other WFS capable browsers.

3.2 Supporting tourism and local tourist industry

SDI data is an important resource for many decisions. Topographic data constitutes an information resource familiar to a broad range of people and therefore extremely valuable in most cases in providing the necessary 'base data' for a literally unlimited number of applications. In the second use case, we explain how existing SDI topographic data encoded in an RDF store and made accessible using LOD can be aggregated by value-added-resellers in the tourism industry to produce tailored geographic information products for local area visitors.

Solutions based on LOD could remedy the drawbacks of data-centric applications and support a broader range of queries. We propose tourism application built-up in the way that makes it possible retrieving RDF geographic information from various sources, literally a LOD data cornucopia provided by governmental entities, tourism sector entities or social groups as well as volunteers and customers. Such data can be accessed through SPARQL Queries using GeoSPARQL vocabulary and RCC8 operators. A federated query should enable access to other RDF data sources that provide URI information. Each query assembles the retrieved data into a linked data resource for supporting the query. The query result can be then displayed in a regular GIS using the LOD4WFS Adapter. Using applications based on SPARQL, a user could compose a request that accesses multiple data sources. The query results can be linked to services enabling purchase of airlines tickets, maybe municipal public transport tickets (zone or time period tickets), hotel accommodation booking, museum tickets, entertainment sites entrance permissions, restaurant reservations, excursion vouchers and so on.

This LOD approach can support a broad range of queries. In particular it includes spatial location factors, spatial relations between subject of interest and methods of profiling attractions and services. Spatial queries (mutual location of objects, transportation bindings, accessibility) can be resolved by GeoSPARQLtriple stores, embracing finding locations and recognizing of topological relations (RCC8). It is also possible to build customized desktop GIS applications, based on QGIS (http://qgis.org), Kosmo (http://www.opensig.es/) or OpenJump (http://www.openjump.org/) open source tools. This will be possible thanks to the dynamic transformation of LOD resources into standard WFS documents. Network applications could be based on the available tools, such as OpenLayers (http://openlayers.org/) libraries or frameworks such as GeoMajas (http://www.geomajas.org/).

4. Outlook

This paper illustrates the potential of linked open data approaches for SDI-type applications based on semantic web technologies. Beyond this rather brief and conceptual overview, a number of issues remain for future work to consider: using SPARQL queries to create new objects; adding metadata generated automatically during queries;creating a reference implementation;assessing the use of ontologies to enrich semantic web application functionality;considering processes for integrating open data that address accuracy and quality concerns;exploring extensions to support real-time sensor data integration;supporting web-based data analytics operations.

Extending the proposed approach to new SDI organizations, the use of dereferable URIs supports more flexible queries. Every spatial data published as LOD (e.g. layers and spatial features) should have unique identifiers, which would allow users to separately acquire and use it in their applications. To emphasize the data hierarchy, the construction of URIs for spatial feature should also contain information about layer to which feature belongs. After URI dereferencing, the results should be available as a RDF document and, for using it in GIS systems, as an on-demand WFS layer created via the LOD4WFS Adapter. In this case the URI of a resources serves also as WFS service interface address.

While this paper only provides some small and future steps, we believe that they contribute to reaching larger goals of the Information Society that Geoinformatics can contribute to.

Acknowledgement

The research was partially supported by a project funded by the National Science Center granted on the basis of the decision DEC-2012/05/B/H/HS4/04197.

References

[1] Battle, Robert, and Kolas, Dave. 2011. Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL Semantic Web Journal.

[2] Berners-Lee, Tim, James Hendler, and OraLassila. 2001. The Semantic Web. *Scientific American* 501.

[3] Bizer, Christian, Heath, Tom and Berners-Lee, Tim. 2009. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS). January 2009.

[4] Chrisman, N. R. 1999. What does ‘GIS’ mean? Transactions in GIS, 3(2), 175-186.

[5] Cohn Anthony G., Bennett, Brandon, Gooday John, GottsMicholas Mark 1997. Qualitative Spatial

Representation and Reasoning with the Region Connection Calculus. *GeoInformatica*, 1, 275–316.

[6] de Man, W. H. Erik. 2006. Understanding SDI: Complexity and Institionalization. *International Journal of Geographical Information Science* 20 (3) : 329-43.

[7] Georgiadou, Yola, S. K. Puri, and S. Sahay. 2005. Towards a Potential Research Agenda to Guide the Implementation of Spatial Data Infrastructures: A Case Study From India. *International Journal of Geographical Information Science* 19 (10) : 1113-30.

[8] Goodwin, J., Dolbear, C. and Hart, G. 2008. Geographical Linked Data: The Administrative Geography of Great Britain on the Semantic Web. Transactions in GIS, 12: 19-30.doi: 10.1111/j.1467-9671.2008.01133.x

[9] Harvey, F., Kuhn, W., Bishr, Y., Pundt, H., &Riedemann, C. 1999. Semantic Interoperability: A Central Issue for Sharing Geographic Information. *Annals of Regional Science*, 33(2), 213-232.

[10] Harvey, Francis, Adam Iwaniak, Serena Coetzee, and Antony K. Cooper. 2012. *Sdi Past, Present and Future: a Review and Status Assessment*. In *Spatial Enabling Government, Industry and Citizens*, edited by Abbas Rajabifard, and David Coleman. Needham, MA: GSDI Association Press.

[11] Hjelmager J., Moellering H., Cooper A., Delgado T., Rajabifard A., Rapant P., Danko D., Huet M., Laurent D., Aalders H., Iwaniak A., Abad P., Daren U., Martynenko A. 2008. An initial formal model for spatial data infrastructures International Journal of Geographical Information Science, Vol. 22 No. 11-12 (Nov. 2008), pp. 1295-1309

[12] Janowicz, K.; Schade, S.; Bröring, A.; Kessler, C.; Maué, P. &Stasch, C. 2010. Semantic Enablement for Spatial Data Infrastructures, *Transactions in GIS* 14 (2) , 111-129.

[13] Jones, Jim, Kuhn, Werner, Keßler, Carsten, Scheider, Simon. Making the Web of Data Available via Web Feature Services. AGILE 2014, Castellón, Spain.17th AGILE Conference on Geographic Information Science, 2014.

[14] Kessler Carsten, KauppinenTomi, Linked Open Data University of Münster. Infrastructure and Applications, "9th Extended Semantic Web Conference (ESWC2012)", 2012, http://data.uni-muenster.de/context/cris/publication/75876

[15] Kuhn, W. (2001). Ontologies in support of activities in geographical space.*International Journal of Geographic Information Science*, 15, 613-631.

[16] Kuhn, W. (2003). Semantic reference systems.*International Journal of Geographic Information Science*, 17(5), 404-409.

[17] Kyzirakos, K., Karpathiotakis M., and Koubarakis, M. 2012.Strabon: A Semantic Geospatial DBMS. In the 11th International Semantic Web Conference (ISWC 2012), Boston, USA, 11-15 November 2012

[18] Marshall, M. Scott and Boyce, Richard and Deus, Helena F. and Zhao, Jun and Willighagen, Egon L. and Samwald, Matthias and Pichler, Elgar and Hajagos, Janos and Prud'Hommeaux, Eric and Stephens, Susie, 2012. Emerging practices for mapping and linking life sciences data using RDF: A case series, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 14, July 2012, Pages 2-13, ISSN 1570-8268, http://dx.doi.org/10.1016/j.websem.2012.02.003.

[19] Masser, I. 1999. All shapes and sizes: the first generation of national spatial data infrastructures. IJGIS, 13(1), 67-84.

[20] Nebert, D. (ed.). 2001. The SdiCookbook, Version 1.1. 2003 :

[21] NeoGeo Vocabulary. Retrieved from: http://geovocab.org/doc/neogeo.html. (Accessed 2014, February, 21th)

[22] Onsrud, H., & Rushton, G. 1995. Sharing Geographic Information: An Introduction. In H. Onsrud& G. Rushton (Eds.), Sharing Geographic Information (pp. xiii-xviii). New Brunswick, NJ: Center for Urban Policy Research.

[23] Rautenbach V., Coetzee S., Iwaniak A. 2013. Orchestrating OGC web services to produce thematic maps in a spatial information infrastructure.Computers Environment and Urban Systems, Vol. 37, pp. 107-120

[24] Shaon, A, Woolf, A, Boczek, R, Rogers, W & Jackson, M. 2011. 'An Open Source Linked Data Framework for Publishing Environmental Data under the UK Location Strategy'. in R Grutter, D Kolas, M Koubarakis& D Pfoser (eds), Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web. vol. 798, CEUR Workshop Proceedings, Terra Cognita 2011 Workshop on Foundations, Technologies and Applications of the Geospatial Web, Bonn, Germany, 23-23 October.

[25] Schade S. and Smits, P. 2012. Why linked data should not lead to next generation SDI. IGARSS 2012: 2894-2897. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06350721. Accessed 6Mar2014

[26] Tosta, Nancy. 1999. *NSDI Was Supposed to be a Verb*. In *Innovations in GIS 6*, edited by B. Gittings. London: Taylor and Francis.

[27] TripleGeo: An open-source tool for extracting geospatial features into RDF triples. Institute for the Management of Information Systems at Athena Research Center. (Accessed 2014, February, 2nd). Retrieved from: https://web.imis.athena-innovation.gr/redmine/projects/geoknow_public/wiki/TripleGeo

[28] Usery, E Lynn, and Dalia Varanka. 2012. Design and Development of Linked Data From the National Map. *Semantic Web* 3 (4) : 371-84.

[29] van Loenen, Bastiaan, Jaap Besemer, and Jaap Zevenbergen. 2009. *Spatial Data Infrastructure Convergence*.In *Sdi Convergence. Research, Emerging Trends, and Critical Assessment*, editedby Bastiaan van Loenen, Jaap Besemer, and Jaap Zevenbergen. Delft, The Netherlands: Nederlandse Commissie voor Geodesie.

[30] Vretanos, Panagiotis A. 2003. Web Feature Service Implementation Specification, Version 1.0.0. Open Geospatial Consortium.

[31] Williamson, Ian, Abbas Rajabifard, and Andrew Binns. 2006. Challenges and Issues for SDI Development. *International Journal of Spatial Data Infrastructures Research*,1 : 24-35.

Session:
Urban Dimension

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Orchestrating the spatial planning process: from Business Process Management to 2nd generation Planning Support Systems

Michele Campagna
Università di Cagliari DICAAR
Via Marengo 2
Cagliari, Italy
campagna@unica.it

Konstatin Ivanov
Tomsk Polytechnic University
Lenina Avenue, 30
Tomsk, Russia
konstantin.Ivn@gmail.com

Pierangelo Massa
Università di Cagliari/ DICAAR
Via Marengo 2
Cagliari, Italy
pmassa@unica.it

Abstract

Metaplanning can be considered as a necessary step for improving collaboration, transparency and accountability in sustainable and democratic spatial decision-making process. This paper reports current findings on the operational implementation of the metaplanning concept developed by the authors relying on Business Process Management methods and techniques. Two solutions are presented which implement spatial planning process workflows thanks to the development of original spatial data and processing services connectors to a Business Process Management suite. These results can be considered as a first step towards the development of 2nd generation Planning Support Systems.

Keywords: Spatial Planning, Metaplanning, Geodesign, Business Process Management BPM, Spatial Web Service, Planning Support Systems

1 Spatial planning, geodesign, and metaplanning

Metaplanning can be defined as the design of the planning process. In real-world spatial planning practices (i.e. Regional Planning or Local Land Use Planning) often metaplanning, as something which is usually not explicitly required by law, is neglected. In such cases taming complex multi-actor planning processes and procedures may result confusing. While on the one hand lack of common understanding among the actors may easily arise, implying difficulties in collaboration, on the other hand understanding how, why, when, by whom planning decisions are made may results blurred both to internal and external stake-holders and observers. The latter should be not considered a minor pitfall as both propositions from advances in planning theory (i.e. Innes' communicative planning, in Khakee, 1998, p. 370) as well as binding regulations on Strategic Environmental Assessment (SEA, Directive 2001/42/EC) –the environmental impact assessment of plans and programmes– require in plan-making not only the evaluation, explanation and documentation of the product (i.e. the final plan) but also of the process. However, what SEA regulations and good practice guidelines usually suggest is the *ex-post* evaluation of some specific part of the SEA-planning process (i.e. degree of public participation in consultation or reliability of data sources), and an *ex-ante* metaplanning approach is most of the time disregarded.

An emerging trans-disciplinary debate among spatial planning and Geographic Information Science scholars concerns the definition and the implementation of the concept of Geodesign [7]. Geodesign can be defined as an integrated process informed by environmental sustainability appraisal which includes project conceptualization, analysis, projection and forecasting, diagnosis, alternative design, impact simulation and assessment, and which involves a number of technical, political and social actors in collaborative decision-making. The innovation in Geodesign, compared to older approaches in environmental planning and landscape architecture, is rather on the extensive use of digital spatial data, processing, and communication resources.

As a matter of facts nowadays, the Information Society reached a mature age, and we face unprecedented wealth in terms of digital (spatial) data sources. The concept of Digital Earth [3] is slowly shaping into reality, and both authoritative and volunteered geographic information resources are available to support analysis and decision-making. Nevertheless in spatial planning, professionals and decision-makers still lag-behind in the digital uptake in the practice, and in properly taking advantage of developing Spatial Data Infrastructures. Hence, making the Geodesign concept operational may be still considered a challenging task.

A small but active research community worldwide, as extensively reported in [6], tried to address these difficulties proposing advanced Planning Support Systems (PSS). By their early proposition [8] PSS were defined as "architecture(s) coupling a range of computer-based methods and models into an integrated system for supporting the planning functions" or more operationally user-friendly microcomputer-based planning system(s), which integrates GIS, sketch tools and spatial models". Indeed, since their early definition PSS were thought as architectures featuring several of the components a Geodesign support system would have. More recent propositions define PSS "a combination of planning-related theory, data, information, knowledge, methods and instruments that take the form of an integrated framework with a shared graphical user interface" [6]. However, it has been noted that the evident obstacles to PPS adoption may be inherent in the concept that comprise first generation PSS [11]. As a matter of facts, most recent perspectives addressing the gap between PSS and real-world urban and regional planning practices concern transparency, flexibility and simplicity [14].

The relevance of the concept of metaplanning, as the activity of specifying actors, activities, methods, tools, inputs and outputs, workflows or in other words the *ex-ante/in-itinere* adaptive design of planning process is also central to the Steinitz's Geodesign framework [13], where the planner (i.e. the coordinator of the Geodesign team) chooses and clearly defines the methods for the study according to a decision-driven approach (i.e. the second iteration), before the resulting workflow is actually implemented (i.e. the third iteration).

According to these considerations, the operational implementation of the concept of metaplanning can be achieved through the description of the planning process. Several attempts have been proposed by scholars to formalise the description of the planning process for diverse purposes, however these results appears to have affected neither the planning practice nor Planning Support System design [7, 1, 6]. As a matter of facts, limitations in Planning Support Systems diffusion may be addressed to lack of flexibility, thus of adaptability to contextual planning process settings.

To address these issues a possible approach is to rely on recent advances in Business Process Management (BPM) [15]. Process-orientation has gained big momentum in the last decade, and BPM techniques and tools have been developed aiming at two main objective: improving process management and easing information system development. BPM found extensive application in industry where goods and services production processes are constantly running and under improvement. Introducing BPM in the production life-cycle requires effort, but it is usually acknowledged that the costs then pay off in the long run as the number of process instances grows.

The authors argue in this paper that PSS design should also be process-driven, rather than technology-driven, and since metaplanning concerns the design and formalisation of the actual planning process, metaplanning should also inform the design of the information systems for planning support. To address this challenge, Business Process Management methods and tools have been applied by the authors to implement the metaplanning concept in the urban and regional planning and Strategic Environmental Assessment domain, claiming that metaplanning may both improve the process and ease customised PSS development accordingly: together the latter results entail the concept of 2nd generation PSS. In this paper the authors report the ongoing results of their research and present original software developed as proof-of-concept of 2nd generation PSS.

2 Implementing metaplanning with Business Process management

The evolution of contemporary spatial governance makes urban and regional planning complex processes -involving actors, activities, resources, objectives, and outputs- which are often difficult to manage in a logical, transparent and accountable manner. As a matter of facts a new figure of planner is emerging as a 'process manager' [16] whose role is the coordination of interacting actors in complex workflows of activities.

Moreover, communication among stake-holders and the broader public is a major issue in SEA, and it can be only correctly realised if proper (i.e. understandable by all) information is given to all the participants [12]: this need also includes information about the process which should explain clearly how, why, and by whom decisions are made. To address these issues a metaplanning approach is proposed by the authors.

Metaplanning can be defined as the explicit design of a (urban and regional) planning process. According to Emshoff [5] poor results of planning are often actually due to poor metaplanning. Since the '70, the concept of metaplanning has been dealt with by several disciplines including artificial intelligence and management science, but it has barely attracted the attention of the planning scholars. As a noteworthy exception de Bettencourt et Al [4] argued metaplanning should be a well-defined step in the plan-making process in order to enhance understanding and coordination among the actors and to achieve expected outcomes. To these Campagna [2] added the enhancement of responsibility, transparency and accountability in the planning process, as well as the definition of the requirements for and the ease of the implementation of process-oriented Planning Support Systems. In order to achieve the latter objectives, Business Process Management (BPM) is proposed in this paper as methodological and technical approach for metaplanning operational implementation.

BPM includes concepts, methods and techniques to support the design and analysis as well as the administration, the configuration, the enactment of business processes [15]. Hence, two are the main objectives of BPM: on the one hand BPM should support the improvement of a process (i.e. business perspective: design and analysis), while on the other hand it should ease the implementation of the supporting information system (i.e. IT perspective: configuration and enactment).

The last decade faced the diffusion of a growing number of software system - Business Process Management Systems (BPMS) - which enact a business process on the base of an explicit process model representation. A Business Process Model (BPm) is a set of activities models and execution constraints among them. From this perspective, urban and regional planning processes can be considered as business processes and Planning Process Models (PPM) can be drawn for descriptive (i.e. as-is) or prescriptive (i.e. to-be) purposes. In planning theory and practice several languages have been used to describe planning processes ranging from natural language descriptions, such as articles in planning regulations, to graphical notations, such as workflow diagrams in planning handbooks. However, most of the latter lack the semantic richness necessary to define planning process models to be used to administrate and enact process instances.

In the last decade, Business Process Model and Notation (BPMN) has been developed and maintained by the Object Management Group as a standard graphical notation for representing business processes in form of diagrams. The rich semantic of this language allows representing actors (i.e. pool and lanes) and activities (i.e. tasks or sub-process) and a variety of executions constraints. Tasks can be manual, automatic or mixed, representing possible diverse situations found in real-world processes: automatic and mixed tasks are those which are supported by the execution of distributed data or processing services. BPMN diagrams are easy to

understand from both humans and machines, becoming the core of business process life-cycle. In facts, many off-the-shelf BPMS feature a BPMN diagram editor for design and analysis, a repository where models are collected, and a process engine which orchestrates the integrated execution of services supporting tasks.
In the reminder of this paper, two examples are presented as proof-of-concepts, aiming at demonstrating the core of this approach on the base of which 2nd generation Planning Support Systems can be implemented.

3 From business process management to Geodesign

The concepts and assumptions presented in the earlier sections have been implemented by the authors in a research project aiming at finding operational way to support both metaplanning and the PSS development from the early stages of the planning process according to a Geodesign approach.

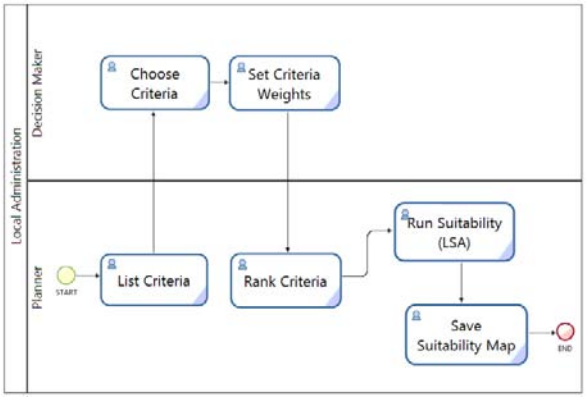
Central to this proof-of-concept is the idea to model the planning process using a BPMN editor in a BPMS and to use the model to orchestrate the technology integration for planning support.
In this project Bonita BPM v6.2.2. suite (referred also to as 'the BPMS' in the reminder) was chosen for it is an open source platform and includes a wide array of BPM functions accessible through a user friendly interface. This BPMS enables the configuration phase of BPM through connectors, which supply functionality (i.e. IT services) to the activities (i.e. the model tasks) by integrating applications, data and services. In the current version connectors to the most used productivity applications and services including email systems, database management systems, information systems (e.g. CRM, ERP, or CMS), web services (using SOAP protocol) are available. For example, business process tasks can send a pre-defined customized email to the customer using an email connector. Unfortunately, no connector is given for accessing spatial data (e.g. WMS, WFS or WCS) and processing services (i.e. WPS). Hence, the first challenge to be addressed in order to implement a test-bed for the implementation of BPM-based metaplanning and for a 2nd generation PSS platform implementation was to create spatial data and processing services connectors for the BPMS.

Two different approaches have been tested so far in the project, including both complex (i.e. online or desktop applications) and atomic components (i.e. spatial data and processing web services). In the next sections two examples are presented, each of which implementing one of the two solutions respectively. The examples are based on a single case study simulating a land suitability analysis (LSA) [9], which can be thought of as a sub-process of a more complex PPM. The LSA sub-process proposed here should be considered as a dummy for the demonstration of capabilities offered by BPM-based approach to planning process design and enactment. This sub-process aims at finding suitable areas for a given land-use according to several criteria. The sub-process entails a number of tasks that should be performed in coordination by different actors in the organizational environment (i.e. the planner and the decision-maker in this example).

The execution ordering of activities and the sequence flow among actors, representing the handover of tasks, can be finely modeled through BPMN in Business Process Diagrams (BPDs). The BPD of the LSA case study is shown in Figure 1.

As shown in Figure 1, in this scenario the planner (P) who is in charge of starting this technical activity (i.e. the LSA sub-process) sets a list of criteria, which is sent to the decision-maker (DM).

Figure 1: Suitability analysis BPD. Model in BPMN.



The DM chooses relevant criteria and then sets weight expressing their relevant importance, and send back the results to P. P ranks criteria values along a suitability scale through a utility function and the runs the analysis calculations. The results of the calculations are then saved.
In the following paragraph this scenario is implemented in two alternative ways.

3.1 Integrating BPMS and GIS

The first solution, provided to orchestrate the technology integration, concerns the call of pre-configured desktop GIS projects from the BPMS during the workflow execution.

For this purpose a custom connector has been developed by the authors taking advantage of the features offered by Bonita BPM. The suite offers several opportunities for the integration of external programs and technologies directly in the workflow through ad-hoc connectors. Connectors can be added to tasks (activities) for accessing external information systems, taking input from the end-user or directly from the process. Bonita BPM offers ready-to-use predefined connectors for several systems and applications and also allows the creation of new connectors from scratch. The connector to call desktop GIS projects during the workflow run has been developed as a system script that allows executing desktop GIS applications in the end-user platform relying on the Windows command shell engine. This capability offered by connectors allows the coordination of work among people and the assignment of specified activities according to individual roles. In the case study example the connector is used to automatically call a pre-configured GIS project in the planner platform to execute the LSA.

Similar GIS workflow management solutions are already available in the market, however in our case unlike in others to our knowledge the control of the workflow execution is performed thanks to the BPD represented in standard BPMN.

In this case, the LSA BPD in Figure 1 is adapted to the technical solution chosen for implementation. In Figure 2a the LSA BPD is shown grouping the activities that are performed by the GIS desktop application by the dashed line, while in Figure 2b the adapted LSA BPD is shown, where grouped tasks are executed within the GIS thus hidden in the diagram.

Figure 2a: Original LSA BPD grouping the activities performed by the GIS desktop application.

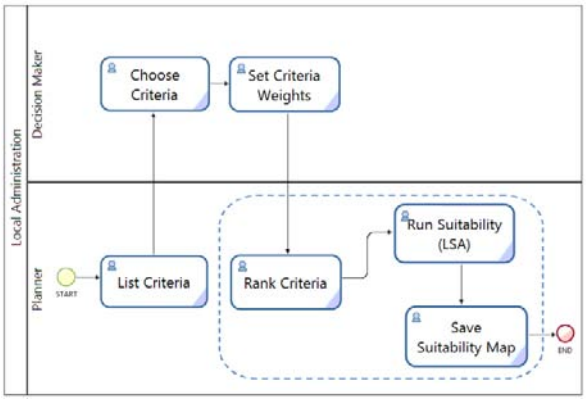
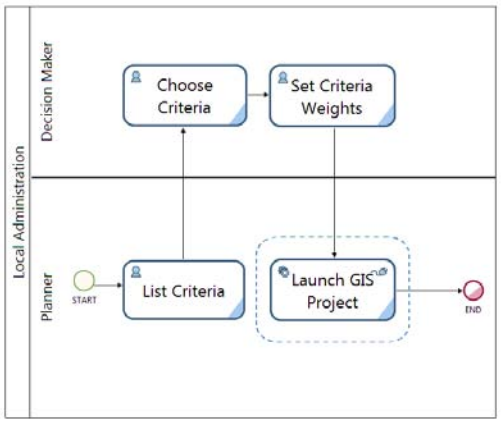


Figure 2b: Adapted LSA BPD in solution 1 (relying on the GIS desktop application connector).



The adapted LSA sub-process is started by P who lists a set of criteria and passes them to DM via a web form. The second activity is performed by DM that accesses the form and chooses criteria. The form template can be designed and implemented directly in the BPMS, offering an input user-friendly interface. After the selection of criteria, when the third activity is activated the platform provides another form, where weights are assigned to criteria according to their relative importance to the DM. The last activity performs the collection of input data, and thanks to the connector the automatic execution of a predefined GIS project in P's workstation. The last part of the process involves the run of the land suitability analysis by P according to DM's input.

The use of a predefined desktop GIS project allows P to perform analysis by means of advanced features offered by GIS applications. In other words, the LSA requires the integration of spatial analytical tools that are supplied in this use case by desktop GIS application. We tested this use-case with both commercial and open source desktop GIS applications. This may be of advantage in urban and regional planning settings for custom GIS project can be prepared by specialists for other professionals.

This first example aims at demonstrating how the integration of BPMS and desktop GIS application offers a technical environment able to coordinate collaborative activities among the actors of a planning process, supplying GIS (and not-GIS) functionalities to the BPMS run-time during the workflow execution. This first solution can be considered viable for planning support in those cases where the task requires relevant flexible human intervention. However, in a number of tasks which may be instantiated in an urban and regional planning process, more advanced automation may improve efficiency. In the next paragraph, a second demonstrator is presented aiming at showing advanced spatial data and services BPMS orchestration possibility.

3.2 Orchestrating WPS by BPMS

The second solution concerns the atomic orchestration of standard spatial data and web services directly within the BPMS. To this end, a custom connector invoking spatial web services (i.e. WFS, WPS) has been developed in Java using Bonita BPM Engine APIs, in order to enable the spatial data and services chaining by the BPMS.

The development of the connector included two steps:

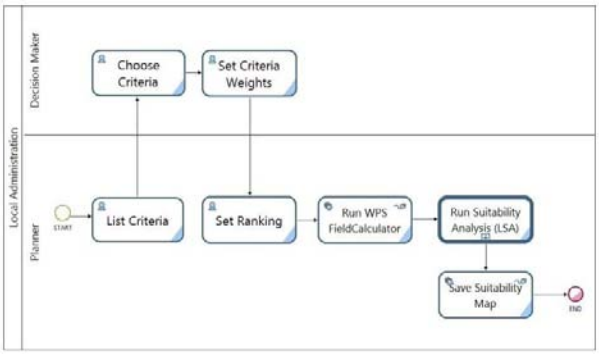
- the connector definition: it controls the external interfaces of the connector (the inputs and outputs), both visible to the users and to the BPMS;
- the connector implementation: where configuration and execution of the connector are defined by implementing default Java class for connectors.

The developed connector requires the user to specify the following parameters: i) a URL of WPS and operation to be executed, ii) input data (e.g. link to WFS and selected features, or input parameters), and iii) the output format (e.g. GML, KML, or shape-file). During the business process execution the connector retrieves and validates input parameters; then it generates xml-encoded request to WPS, containing input parameters (e.g. WFS link and features, processing operation). This request is then submitted to the URL of the WPS. The WPS performs the request querying data from WFS and processing input data (including input parameters), and returns xml-encoded response to the connector. The connector receives this response and saves the results into the global variable of the business process.

Figure 3 shows the adaption of the base LSA BPD to the spatial web services orchestration solution. The first three activities (list criteria, choose criteria and set criteria weights) are performed by humans, hence they are the same as in the previous solution. The fourth activity is performed by a planner who sets ranks manually in this example. The next activity reads stored ranking data, acquires input layers as WFS features and parameters for WPS execution, then requests the WPS to run the thematic attribute 'field calculator' process. In this experiment we used the 52°North WPS with 220+ SEXTANTE Processes extension on Apache Tomcat 7.0. The result of the execution is then transmitted to the sub-process which invokes a WPS operation for the criterion map 'Union' and eventually the WPS executes the field calculation which performs the weighted sum. The last activity takes the result of the LSA and saves the output

suitability map in the location specified by the user thanks to another simple connector developed by the authors. The saved suitability map can be opened in a desktop GIS application or published as WMS or WFS. The later step is currently under development, thus it is not included in the model in figure 3.

Figure 3: Suitability analysis BPD with the introduction of the connector for spatial web services.



The purpose of this second case study is to demonstrate the orchestration of spatial web services via BPMS. Unlike the previous example, in this case a greater programming effort was required. However, this second solution may open further alleys for 2nd generation PSS development for it enables a higher level of computer support to humans thanks to the orchestration.

4 Conclusions

Recent advances in urban and regional planning, enhanced complexity in spatial governance, and Strategic Environmental Assessment call planners to novel approaches to planning process management and assessment. The authors propose the concept of metaplanning as viable solution for planning process improvement in term of actor collaboration, and process transparency and accountability. Accordingly a novel BPM approach to metaplanning is proposed.

The authors claim that a BPM approach to metaplanning may also ease the agile development of process-oriented 2nd generation Planning Support Systems. To proof this concept alternative technology solutions are proposed which demonstrate with reference to a simple process metaplanning in action.

The early results of this research project can be considered as a first contribution towards the creation of an architectural framework for 2nd generation Planning Support System design and implementation.

Acknowledgements

The work presented in this paper was developed by the authors within the research project “Efficacia ed efficienza della governance paesaggistica e territoriale in Sardegna: il ruolo della VAS e delle IDT” [Efficacy and efficiency of landscape and environmental management in Sardinia: the role of SEA and of SDI] CUP: J81J11001420007 funded by the Autonomous Region of Sardinia under the Regional Law n° 7/2007 "Promozione della ricerca scientifica e dell'innovazione tecnologica in Sardegna".

References

[1] R. Brail and R. Klosterman. *Planning Support Systems: integrating Geographic Information Systems, models, and visualization tools*. ESRI Press, Redlands, CA, United States, 2001.

[2] M. Campagna. Geodesign, Planning Support Systems and Metaplanning, *Disegnare con*, 6(11):133-140, 2013.

[3] M. Craglia, K. de Bie, D. Jackson, M. Pesaresi, G. Remetey-Fülöpp, C. Wang, A. Annoni, L. Bian, F. Campbell, M. Ehlers, J. van Genderen, M. Goodchild, H. Guo, A. Lewis, R. Simpson, A. Skidmore, P. Woodgate. Digital Earth 2020: towards the vision for the next decade. *Intl. Journal of Digital Earth*, 5(1), 2012.

[4] J.S. de Bettencourt, M. B. Mandell, S. E. Polzin, S. L. Sauter, J. L. Schofer. Making planning more responsive to its users: the concept of metaplanning. *Environment and Planning A*, 14(3):311-322, 1982.

[5] J.R. Emshoff. Planning the process of improving the planning process: A case study in meta-planning. *Management Science* 24(11):1095-1108, 1978.

[6] S. Geertman and J. Stillwell. Planning support systems: content, issues and trends. In S. Geertman and J. Stillwell, editors, *Planning support systems best practice and new methods*, pages 1-26. Springer, Dordrecht, The Netherlands, 2009.

[7] M. Goodchild. Towards GeoDesign: Repurposing cartography and GIS? *Cartographic Perspectives* 66: 7–22, 2010

[8] B. Harris. Beyond Geographic Information Systems: computer and the planning professionals. *Journal of American Planning Association*, 55(1):85-90, 1989.

[9] L. Hopkins. Methods for generating land suitability maps: a comparative evaluation. *Journal for American Institute of Planners*, 34(1):19-29, 1977.

[10] A. Khakee. Evaluation and planning: inseparable concepts. *Town Planning Review*, 69(4):359-374, 1998.

[11] R. Klosterman. Preface. In S. Geertman and J. Stillwell, editors, *Planning support systems best practice and new methods*. Springer, Dordrecht, The Netherlands, 2009.

[12] M. Partidario. *Strategic Environmental Assessment Better Practice Guide - methodological guidance for strategic thinking in SEA*. Governo de Portugal, available at http://www.iaia.org/publicdocuments/special-publications/SEA%20Guidance%20Portugal.pdf [last visited 01.03.2014].

[13] C. Steinitz. *A framework for Geodesign*. ESRI Press, Redlands, CA, United States, 2012.

[14] M. te Brömmelstroet. Transparency, flexibility, simplicity: From buzzwords to strategies for real PSS improvement. *Computers, Environment and Urban Systems*, 36(1):96–104, 2012.

[15] M. Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer-Verlag, Berlin Heidelberg, 2012.

[16] B. Zanon. Planners' Technical Expertise: Changing Paradigms and Practices in the Italian Experience. *Planning Practice & Research*, 29(1):75-95, 2014.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Recitoire: a tool for qualitative surveys involving citizens in urban planning projects

David Noël
Steamer Research Team
Grenoble Computer Science Lab
681 rue de la Passerelle, 38400 Saint-Martin d'Hères, France
david.noel@imag.fr

Marlène Villanova-Oliver
Steamer Research Team
Grenoble Computer Science Lab
681 rue de la Passerelle, 38400 Saint-Martin d'Hères, France
marlene.villanova-oliver@imag.fr

Jérôme Gensel
Steamer Research Team
Grenoble Computer Science Lab
681 rue de la Passerelle, 38400 Saint-Martin d'Hères, France
jerome.gensel@imag.fr

Abstract

The difficulty to involve citizens into projects that influence or transform their experience of the urban space is underlined by public authorities and professionals such as urban planners. The implication of citizens in the existing modes of consultation (public meeting, opinions polls) is often limited and not representative. New solutions for facilitating citizens' involvement in both the diagnosis and the construction of the city need to be found. We propose here a prototype called Recitoire, as a support for qualitative surveys involving citizens in urban planning projects. Using a mobile application, a data collect is performed which includes the path followed by a citizen (her/his trace is kept) and the different kinds of media files she/he produces all along the path to illustrate her/his feelings and impressions on a given thematic - chosen for the survey by urban planners. A server application centralizes the collected data and offers an interface for both their exploration and their exploitation by the actors of the urban project.

Context

Amplified by the advent of Google Maps in 2005, an interactive cyber-mapping has developed strongly in which contributors describe themselves the space where they live, giving a contemporary form to geography, called neogeography [2]. This new behaviour questions directly public authorities, professionals working for the town and country planning, and scientific communities about the evolution of tools and the role of the different actors involved in this field.

In particular, the evolution and practices of new technologies linked with spatial data lead to new applications, which might deeply change the way to think urban and territorial project [1, 2]. These new solutions include giving the voice to the urban space users – also users of the local amenities offered on the territory –, considering "citizens as sensors" according to the expression of Goodchild [2].

In France, consultation practices applied and followed by project managers respect the rules defined in the law related to solidarity and urban reshaping (December 2000) and by the Town Planning Code (for example public consultations and meetings). However, research in territorial sciences has suggested going beyond such practices and has shown the importance of an effective and continuous communication between all the actors who could feel concerned by the design, development and management of urban and territorial projects. Researchers also confirmed the hypothesis that data collected by citizens improve the information used in the decision making process in the domain of spatial planning [3]. Furthermore, the integration of Volunteered Geographic Information into urban management program is identified in recent work [4] as a research challenge, especially for the definition of useful types of contribution in urban management.

In this context, our research group called FabTer[1] ("Fablab Territoires") has initiated some work around new methodologies and associated software solutions to encourage citizens' implication in the process of making a diagnosis about the territory and addressing urban planning issues. We have designed, prototyped and tested a client/server application called Recitoire that aims at collecting qualitative data from citizens and at visualizing and analysing them through a specific interface for urban planners. The name Recitoire comes from the contraction of the French words *Récit* (narrative) and *Territoire* (territory), and refers to the narrative setting of the territory that is expected from the users. Using a mobile application, a data collect is performed which takes the form of both a path followed by the citizen (the track is kept) and the media files she/he produces all along the path to illustrate her/his feelings and impressions on a given thematic. A server application centralizes the collected data and offers an interface for both their exploration and exploitation. Recitoire therefore serves as a support for conducting surveys with the citizens concerned by a given urban project.

In this paper, we first present our motivations in designing the application Recitoire. Then, we explain how a survey is managed with the help of both the mobile and server side of the application. A review of related works is presented. We give some elements regarding the first tests we have conducted before we conclude and discuss further works.

2 Motivations

[1] The collaborative research project FabTer is made up of researchers of Pacte (UMR 5194) et LIG (UMR 5217, Steamer group) labs and of the consulting firm DéTOUR (all in Grenoble, France).

Citizens are expected to participate actively and from its very early stage to any urban planning project that is supposed to transform their city. We claim that "citizens as sensors" approaches, and associated tools, do have some assets to fulfil this twofold requirement - *actively and from its early stage.* Through Recitoire, we propose a Smartphone application that allows users to contribute in an easy and attractive way to give their opinion about their everyday space (where they live there, where they work, etc.). They can express their own feelings about this space, as it is or as they would like it to become, and document their contributions with various media (video, photo, audio, text).

The particularity of our approach is that the contributor is expected to report a consistent reasoning about her/his urban space instead of an inventory of unrelated observations. For that purpose, we ask the citizen-contributor to tell a story she/he illustrates with images, sounds and texts. We call this story an *urban narrative*. It means that all the contributions should not be completely independent from each other and that it is possible to find the meaning of this urban narrative. The application Recitoire Mobile has been therefore developed to facilitate the emergence of a story, offering thus some support to qualitative surveys about citizens' territory. A citizen can use this mobile application on a path that she/he follows every day (for example to reach her/his workplace) or fixed for the study, having in mind to tell a story on a specific thematic.

The application we propose has been designed and developed to allow urban planners to easily configure a survey according to their needs as presented in the next section.

3 A survey supported by Recitoire

3.1 Overview

Let us suppose that the manager of a urban project wants to conduct a survey with the residents of the impacted sector. Together with urban planners, she/he designs and plans the study as a classical one (when, with who, on which topic, etc.), but having in mind that citizens will use the application Recitoire Mobile for collecting the data related to urban narratives. The mobile application can be configured accordingly using specific functionalities provided by the application Recitoire Server. The data collect is made by and thanks to citizens equipped with Recitoire Mobile on their devices. Data are then centralised through Recitoire Server and available for an analysis by spatial planning expert or project manager.

3.2 Configuration of the Recitoire Mobile Application

The qualitative survey to carry out is characterized by a given theme on which citizens will have to contribute. "Your route to workplace on foot" or "Places for a break during a walk with children in your area" are examples of such themes. A theme is the topic on which contributors will be invited to tell a story when they will start the application. The theme (a short sentence) can be associated with a picture for look and feel purpose. A set of keywords is also defined that will be used by citizens in order to tag their *contributions*. A contribution is composed of a media files (photo, viedo, etc.) localized (in space an time), associated with an optional comment and tagged, all being in relation with the theme of the study. Together, all the contributions made on a route form a urban narrative.

The project manager has then to decide at which (time or space) interval a contributor will be invited to contribute. For example, she/he can impose contributions each 20 meters if she/he want very frequent feedback (which could be adapted for the theme "Your route to workplace on foot") or give no constraint (free contributions are more adapted to the theme "Places for a break during a walk with children in your area" or to the theme "Your route to workplace" if a means of transportation is used).

Finally, the project manager can choose information she/he needs to know about the contributors. This information will form the personal profile of the contributor.

To perform this configuration, the project manager uses the application Recitoire Server (see section 3.4 and 3.5). A succession of forms allows her/him to enter the different parameters (related to theme, contribution interval and user profile). An XML file is generated at the end of the process. This file is then stored on each device and is used to configure the mobile application.

3.3 The Recitoire Mobile Application

Recitoire Mobile has two main functionalities: Create a urban narrative (that is, a data collect mode) and Visualize a urban narrative (that is, localize it on a map and access its contributions). The application main interface for data collection is shown in Figure 1.

A story-databoard is the receptacle of the citizen contributions. Inspired by the cinematographic concept of « storyboard », it shows the contributions and their associated

Figure 1 : Recitoire Mobile (Create a urban narrative)

comments appearing one after the other in the chronological order on the screen using horizontal scrolling. The idea is to provide a visualization of the whole urban narrative on the screen. To contribute, the citizen chooses the media she/he wants to create (photo, video, audio). After recording, she/he is invited to comment her/his contribution in an audio or textual way.

If the contributor is walking, a movement bar appears under the story-databoard indicating the distance and the different names of the locations she/he has passed by all along the route. For a contributor using a means of transportation, the movement bar can be disabled.

In both case, at the beginning of the route, she/he has to give names to her/his starting and arrival points and, at the end of the route, to confirm that she/he is arrived at the ending point. The contributor is also invited to describe her/his route, in order to make easier the future exploitation of her/his story. She/he must therefore comment it, giving it a title and choosing keywords in the predefined keywords list.

The user can visualize the route on a map and she/he can play again the contributions she/he has made (see Figure 2).

Figure 2 : Recitoire Mobile (Visualization of a Narrative)

3.4 The Recitoire Server Application

The Recitoire Server application has been introduced in section 3.2 where the configuration functionality is presented. This application is also dedicated to import and exploit all data collected using the Recitoire Mobile Application.

The citizens' narratives are downloaded from mobile devices and loaded into the Recitoire Server database. For the moment, data extraction requires connecting the mobile phone used to a computer for a direct transfer from Recitoire Mobile to Recitoire Server. This task is easily achievable thanks to a wizard.

Once uploaded, the data produced by the citizen are available in the server application interface for consultation and analysis (Figure 3).

The list of all created narratives within the context of the current study is proposed. It is possible to filter urban experiences according to many criteria: temporal (date, time, journey duration), spatial (particular area of the investigated sectors, distance travelled), thematic (keywords), media type (photo, video, audio, text) and contributor features (components of her/his profile) (see left side on Figure 3).

For each urban narrative, the simultaneous visualisation of the route and of the story built all along is proposed. The story appears according to the story-databoard model (with contributions and comments). The contributions are also located on the map and made visible by clicking on the corresponding point.

The project manager or urban planers can use this interface to explore and analyse the routes and contributions in order to build some report summarizing the contributors urban experience feedback in line with the theme that have been defined for the survey.

3.5 Technical aspects

To date, Recitoire Mobile is designed for Android System only. SQLite is used to store data and the OSMdroid API to display OpenStreetMap Maps.

The server application is developed in PHP, Jquery, HTML5 and CSS3. It also uses the bootstrap CSS framework. Data are stored through SQLite and the OpenLayers API is used to display OpenStreetMap Maps.

Apart from the interfaces and functionalities presented in this paper, Recitoire Server provides users with two wizards. The first one concerns the definition of the survey parameters and the generation of appropriate configuration files for the device mobiles to be used by citizens. The second wizard helps to upload the collected data from the mobile devices to the server application. These functionalities allow the actors of urban projects to be autonomous when using Recitoire.

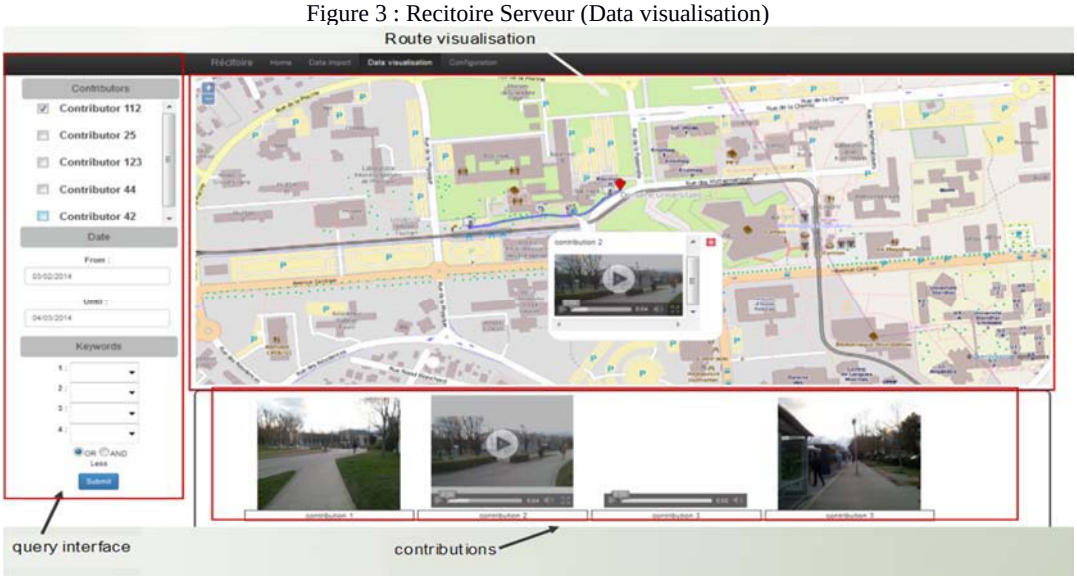
5 Related works

There are many applications or frameworks dedicated to the creation of VGI applications. Mocomapps [5] and Sensr [6] allowed non-programmers to create their own VGI applications. These applications rely on the principle of re-using predefined interface elements. This principle allows non-programmers to create their own applications but it does not fit with the creation of a more complex application like Recitoire.

More closely related to our work, wq [7] is a modular framework for VGI applications. It enables programmers to reuse modular components in order to facilitate the creation of VGI applications. However, in our case, the underlying model we use (a set of multimedia contributions localized possibly at regular interval temporally organised) imposes a very precise data structure that does not fit with the available frameworks.

Many applications are comparable to Recitoire as they allow for a citizen data collection from mobile devices including media contributions. PhotoMap[2] [8] is an application that we

Figure 3 : Recitoire Serveur (Data visualisation)



[2] http://photomap.ingi.ucm.cl/PhotoMap/

have developed and that has inspired our work on Recitoire, especially the combination between the route and geolocated photos along it.

The application Beecitiz[3], used by several French municipalities, or FixMyStreet[4], are based on the principle of FixMyCity[5] : in a specific town, citizens report local problems (degradation, erased marking...) in order to help services of the municipality. This clearly aims at inventorying isolated and negative events but not at constructing a tale providing a global feeling about the surrounding urban space. This application can help one municipality to fix day-to-day problems while the Recitoire application addresses more issues related to the co-construction of the city together with citizens.

The application Mapiness[6] collects data from citizens about their feeling in terms of well-being level but not their feedback about their own urban experiences.

The application MyFunCity[7] can be compared with Recitoire because it also aims at impacting public policy but not in a particular project and from its beginning (diagnosis phase). Indeed, the initiative to use MyFunCity is not configured and launched by a urban project manager: the citizens are simply invited to express themselves about a whole city.

The idea of multimedia citizen contributions is also present in the implementation of contribution platforms based on the LOCAST[8] framework, for example in the project Rio South Mapping[9]. Comparatively, Recitoire is a tool designed for the particular needs of urban planners and whose set up does not require particular computer skills.

6 First tests

We have conducted some user tests with twenty students of the Urbanism Institute of Grenoble (Urban Design Master degree) and twenty citizens encountered during the Experimenta[10] Exhibition. They have been asked to use Recitoire Mobile on a case study, and then to fulfil a questionnaire built in order to assess:

- The degree of satisfaction of users with regard to the principle of the contribution that is expected from them : the construction of an illustrated tale on a thematic imposed.
- Their feelings about the functionalities offered by Recitoire Mobile and their usability.

The principle of telling a story instead of creating independent contributions is considered useful or essential by 70% of the users. Conversely 25% considered it unnecessary. Furthermore, the opportunity to review these stories on the phone is useful or essential for 90% of the users.

We can also notice that the diversity of the media available in order to contribute is widely considered as essential (90%).

The functionalities are reported appropriate and useful for 35% of the users but incomplete by 35% and inadequate by 5%. Some of the users suggested a cartographic visualisation of the area during the route. The use of these functionalities is considered as simple or fairly simple by 75% of the users. However, though 75% of users consider as clear the explanations they were given before they use the application, 35% also suggest a self-training session including examples of what is expected.

7 Conclusion and future work

In this paper, we have presented Recitoire, a client server application that aims at facilitating the implication of citizens in the very first step of the lifecycle of a urban project. Qualitative surveys can be organised using the mobile application. A data collect can be performed involving citizens who are asked to build a narrative about a thematic following a path and illustrating it with the help of media files. A server application centralizes the collected data and allows the actors of the urban project to explore and exploit the feedback of citizens through a specific interface.

Our first tests with the prototype show that citizens welcome such an approach. The application has been set up together with specialists of urban planning but we intend to make other tests with policy makers for instance.

As future work, we address two issues. First, we need to make the approach evolve so that studies at a larger scale (in terms of territory but also of number of respondents) can be led. This requires for instance to develop functionalities for the online upload of data collected by citizens from their personal devices. Second, we intend to propose more advanced functionalities at the Recitoire server side for the processing and the analysis of the data collected. This is particularly important if one considers that the volume of collected data increases.

References

[1] Elwood S. *Geographic Information Science: new geovisualization technologies – emerging questions and linkages with GIScience research.* Progress in Human Geography 33 (2): 256 263., 2009

[2] Goodchild M. F., 2009, *NeoGeography and the nature of geographic expertise.* Journal of LocationBased Service 3(2) : 82 96

[3] C. Seeger C. (2008). *The role of facilitated volunteered geographic information in the landscape planning and site design process.* GeoJournal, 72 (3) : 199-213.

[4] Song W. and Sun G. *The role of mobile volunteered geographic information in urban management. In* 18th International Conference on *Geoinformatics*, pages 1-5, June 2010.

[5] Hupfer S., Muller M., Levy S., Gruen D., Sempere A., Ross S., and Priedhorsky R.. MoCoMapps: *mobile collaborative map-based applications.* In ACM CSCW Companion '12, page 43-44. ACM, 2012.

[6] Kim S. and Paulos E.. *A subscription-based authoring tool for mobile citizen science campaigns.* In ACM CHI Extended Abstracts '12, page 2135-2140. ACM, 2012.

[7] Sheppard, S. A. (2012). *wq: A Modular Framework for Collecting, Storing, and Utilizing Experiential VGI.* ACM SIGSPATIAL GEOCROWD'12 Nov. 6, 2012.

[8] W. Viana, J. Bringel, J. Gensel, M. Villanova-Oliver, H. Martin. *PhotoMap: From Location and Time to Context-Aware Photo Annotations.* J. of Location Based Services, vol 2, pp 211-235, 2008.

3 http://www.beecitiz.com/
4 http://www.fixmystreet.com
5 http://fixmycityapp.com/
6 http://www.mappiness.org.uk/
7 http://myfuncity.uol.com.br/
8 http://locast.mit.edu/
9 http://www.beecitiz.com/
10 http://www.experimenta.fr/

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Planned vs. Real City: 3D GIS for Analyzing the Transformation of Urban Morphology

Pilar Garcia-Almirall
Universitat Politècnica de Catalunya
Diagonal 646
Barcelona, Spain
pilar.garcia-almirall@upc.edu

Francesc Valls Dalmau
Universitat Politècnica de Catalunya
Diagonal 646
Barcelona, Spain
francesc.valls@upc.edu

Montserrat Moix Bergada
Universitat Politècnica de Catalunya
Diagonal 646
Barcelona, Spain
montserrat.moix@upc.edu

Abstract

Cities are constantly evolving: buildings are built and demolished, altering the landscape of our cities; Urban Plans describe what we want our cities to be, undergoing revisions as we change our vision of the future of our cities. This paper presents a methodology to model the interactions between what the city is and what it wants to become. The old quarter of Sant Andreu in Barcelona (Spain) was used in a pilot study for the development of a methodology to automatically quantify and visualize the outcome of regulation changes as a strategic tool for the Urban Planning Department of the Barcelona City Council. This paper describes a methodology developed to measure the magnitude of the buildings conformity or disconformity to the determinations of the Urban Plan (current and proposed), and to display this information in 3D, to allow a more natural interpretation of the results. Special care was put into the methodological approach to ensure that it could be replicated at neighbourhood or city-wide scales. A methodology for the interpretation of the heights of staircase towers and ventilation courtyards from neighbouring entities heights, through the analysis of adjacency relationships in a non-topological Geographic Information System is also discussed.

Keywords: 3D GIS, Urban planning, Urban regeneration, Local management, Smart visualization, Topology

1 Introduction

Urban processes driven by changing social, economic and environmental factors must be rethought [1, 2]. The development of the cities of the future does not necessarily need to be the sprawling growth of past decades but the regeneration of existing urban spaces [3]. Technology can help understand these processes and allow policymakers to make better decisions [4] with deeper public participation [5] for more adaptable, dependable and liveable cities [6].

In the consolidated city and in particular its historic centre, the complexity of the urban pattern makes the regeneration processes (both physical and social) a difficult task; nonetheless, these are usually the areas with a more pressing need for urban renewal processes [7].

Transformation processes in these settings must handle information about the morphology of the built environment with a sensitivity that considers urban quality, sustainability and security, while giving response to the needs of its inhabitants [8].

In areas undergoing regeneration processes, planners need to know with precision both past and present state of the city to plan for a better future. With this objective, a strategic tool was developed to aid in the decision-making process, using the vast pool of urban data stored in cadastral and planning databases to (a) accurately quantify the conformity with height regulations of today's built environment and (b) to evaluate the outcomes of modifications of height regulations.

The object of this paper is to explain the methodology developed to precisely quantify the complex interactions between built reality and urban regulations.

1.1 Case of study

The case of study chosen for the development of the methodology was the old quarter of the Sant Andreu District in Barcelona.

The buildings in the area of study (Figure 1) are part of the centre of the former town of Sant Andreu, from which the district takes its name, and the development it underwent in the 19th century when it was incorporated to Barcelona. This historic development resulted in a complex urban structure suitable to use as workbench to test the methodology, with an area of 90 hectare containing 2,775 parcels, distributed in 148 city blocks around its main commercial street, *Gran de Sant Andreu*.

The area of study was in the process of modifying the planning regulations [9] in effect since 1976 in 27 municipalities of the Barcelona Metropolitan Area. Being a plan from 1976, some parameters such as the maximum height were not drawn explicitly in the planning regulations, and had to be interpreted [10] with the aid of the staff in the Urban Studies Bureau of the Urban Planning Department of the Barcelona City Council.

The case of study was chosen because of the interest the city planners had in knowing with precision the compliance of the buildings in the area with the maximum heights allowed by the planning regulations.

The methodology developed was very valuable to evaluate the possible outcomes of modifications of different parameters of the regulations, and made possible to assess the current built mass of the whole city.

2 Methodology

2.1 Interpretation of the height of ventilation courtyards and staircase towers

The height of buildings was stored as 2.5D cartography in a sub-parcel dataset (being sub-parcels pieces within a parcel with a distinct height from neighbouring sub-parcels). Height information was stored as an alphanumeric string which encoded several pieces of information about the sub-parcel, including the number of floors below and above street level. For example, a sub-parcel with a "–II+V" attribute had two subterranean floors and five floors above street level.

In the case of ventilation courtyards (VC) and staircase towers (ST), sub-parcels did not have a height attribute but a code that identified them as such ("P" for VC and "E" for ST). However, ST volumes protrude from flat roofs and VC are considered by planning regulations as part of the building, and consequently both types needed to be assigned a height value.

Since VC accounted for almost 20% of the area of all sub-parcels and ST for another 2%, to get accurate results a methodology had to be developed to automatically assign height values to this types of entities from their spatial context, considering that ST are as high as the top floor they serve in their parcel and VC have a lightweight roof at the level of the lowest floor they serve in their parcel.

For sub-parcels representing ST, the assigned value was the maximum height of all adjacent sub-parcels belonging to the same parcel (Figure 2) and for sub-parcels representing VC, the assigned value was the lowest height of all adjacent sub-parcels belonging to the same parcel (Figure 3).

The calculation of the height to be assigned to ST and VC involved two topological relationships: (a) adjacency to other polygons but (b) considering only polygons inside the same parcel. Since the Geographic Information System (GIS) used was non-topological, a methodology had to be implemented in Structured Query Language (SQL):

1) A tool to convert lines to polygons was used to get a table with 3 ID fields: line ID, left polygon ID and right polygon ID (*lines*) for all sub-parcels (*volumes*).
2) A dictionary of key-value pairs (*type_dictionary*) was made to translate the alphanumeric encoding to a numeric value (*volumes_height*) measuring the number of floors above street level (Figure 4, above).
3) A table with the attributes of the polygons on both sides of each line was built (*lines_volumes*) from the tables described previously (Figure 4, below).
4) This intermediate table had to be reshaped as a list using union queries, excluding the polygons not pertaining to the parcel the fragment belonged to and excluding the sub-parcels that were not ST (figure 5, left) or VC (Figure 5, right).
5) An aggregation query for each type was performed to get the corresponding values of the height attribute: the highest value of all neighbours for ST and the lowest value of all neighbours for VC.

Figure 1: Area of study in the Sant Andreu District of Barcelona

Figure 2: Correction of staircase towers (red) of two neighbouring parcels (sub-parcels in brown and blue hues)

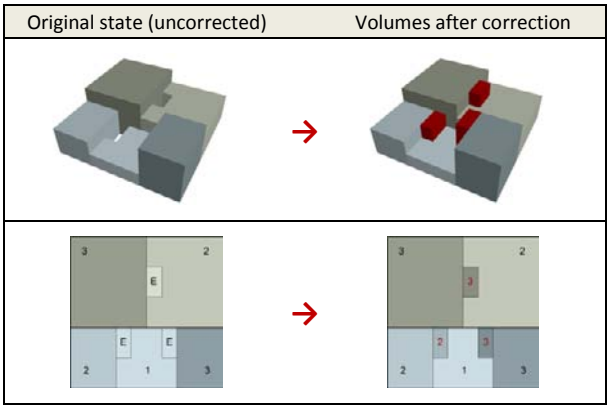
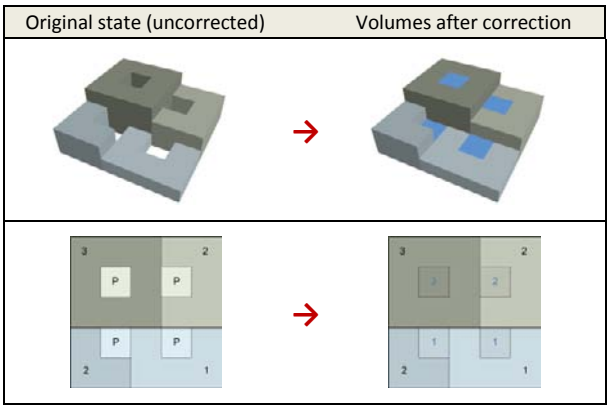


Figure 3: Correction of ventilation courtyards (blue) of two neighbouring parcels (sub-parcels in brown and blue hues)



2.2 Overlay operations

With the sub-parcels volumes corrected and their height converted to a numeric value, their heights were compared to the height the plan allowed using two Boolean spatial operations: (a) a spatial intersection (Figure 6, left), the result of which was the fragments of sub-parcel inside zones and (b) a spatial difference (Figure 6, right), the result of which was the fragments of sub-parcel inside systems (roads and parks).

With the result of the overlay operations it was possible to determine the conformity to the urban plan for each of the fragments (Table 1) from their building height (HB) and planned height (HP).

The magnitude of the conformity (measured in area units) for each resulting sub-parcel fragment was represented in a map, multiplying its area by the corresponding number of floors beneath or exceeding the allowed height (Figure 7). This map was a very valuable analytical tool, but to visualize its information in a more intuitive way a different approach had to be developed to make it easier to interpret.

Table 1: Types of fragments from the overlay operations

Fragment	Plan entity	Condition	Operation	Symbology
Underbuilt	Zones	HB - HP < 0	Intersection	Blue hues
Conformant	Zones	HB = HP	Intersection	Grey
Overbuilt	Zones	HB - HP > 0	Intersection	Pink hues
Overbuilt	Systems	HB > 0	Difference	Dark green
Conformant	Systems	HB = 0	Difference	Light green

2.3 Aggregation at parcel level

It is not legally allowed to compensate overbuilt volumes with underbuilt ones inside a parcel (Figure 8), and accordingly aggregate calculations had to be performed separately for both situations to avoid the aggregate operations adding positive and negative numbers (which would be mathematically correct but not possible according to the regulations).

Formulae 1 to 5 show the aggregation operations to calculate for each parcel: the total built area (1), the maximum allowed built area in zones (2), the overbuilt area in zones (3), the underbuilt area in zones (4), and the overbuilt area in systems (5).

$$Real\ Built\ Area_{Parc} = \sum_{Subp \in Parc} HB_{Subp} \cdot A_{Subp} \quad (1)$$

$$Allowed\ Built\ Area_{Parc} = \sum_{FragZ \in Parc} HP_{FragZ} \cdot A_{FragZ} \quad (2)$$

$$Overbuilt_{Parc} = \sum_{FragZ \in Parc} \left(HB_{FragZ} - HP_{FragZ}\right) \cdot A_{FragZ} \quad (3)$$

$$Underbuilt_{Parc} = \sum_{FragZ \in Parc} \left|HB_{FragZ} - HP_{FragZ}\right| \cdot A_{FragZ} \quad (4)$$

$$Overbuilt\ in\ Systems_{Parc} = \sum_{FragS \in Parc} HB_{FragS} \cdot A_{FragS} \quad (5)$$

Figure 4: Queries to build the neighbours attributes table

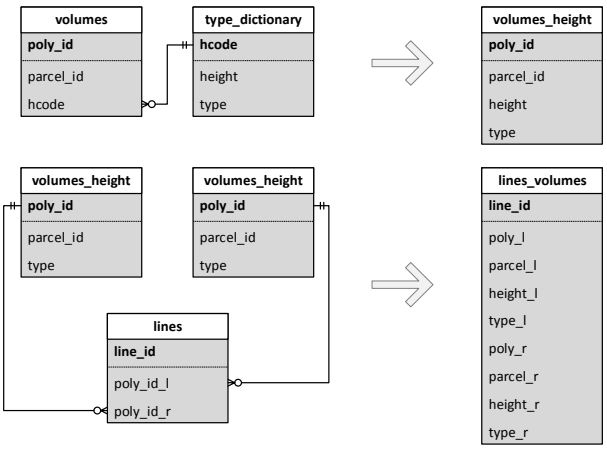


Figure 5: Operations to obtain all neighbouring volumes inside the same parcel for every ST (left) and VC (right)

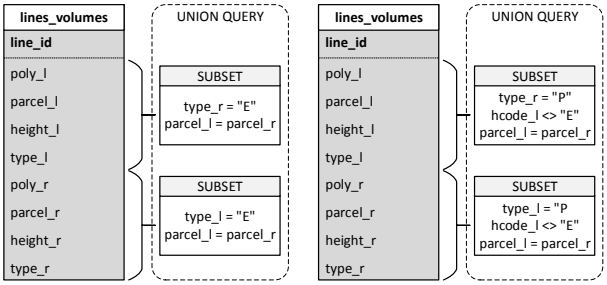


Figure 6: Spatial intersection (left) and difference (right)

Figure 7: Conformity to the urban plan of each fragment

Using the following notation:

- **HB**: Real height of the building, measured in number of floors above street level.
- **HP**: Maximum height according to the plan, measured in number of floors above street level.
- **A**: Area of a polygon entity (surface of land occupied).
- **Parc**: Set of parcels inside the area of study.
- **Subp**: Set of sub-parcels.
- **FragZ**: Set of fragments from the spatial intersection between parcel and planning layers (in zones).
- **FragS**: Set of fragments from the spatial difference between parcel and planning layers (in systems).

3 Results

3.1 Parcel level results

The aggregated values were displayed in choropleth maps, to visualize the magnitudes of overbuilt and underbuilt areas of each parcel. These maps were a strategic tool for the City Planning Department to visualize and identify the parcels with the most outstanding values.

With the map of overbuilt areas of parcels (Figure 9), planners were able to identify the parcels with a higher degree of excess volume and to visualize the spatial clustering of overbuilt parcels facing certain streets or concentrated in specific city blocks.

The map of underbuilt areas in parcels (Figure 10) allowed planners to visualize the places where underbuilt parcels were clustered together as candidates to successfully implement transformation policies.

3.2 A new approach for the representation of fragment level results in 3D

The representation of the results using 2D maps was unable to convey the complex volumetric information successfully because height data had to be abstracted to be represented in plan view as colour scales, hatch densities or labels.

The use of 3D imagery allowed the authors to represent the volumes in a more natural and intuitive way since it matched the way we experience our cities.

Figure 11 shows the criteria to display the overlapping information of overbuilt and underbuilt fragments. The third dimension allowed the authors to display overlapping information without having to resort to 2D representation constructs such as transparency or hatching.

Figure 12 shows the results for the case of study, where the viewer is able to visualize and relate two concepts simultaneously (real height and planned height) much more easily than using 2D maps.

An axonometric aerial photograph was compared to the corresponding result (Figure 13) to highlight the value of the methodology developed as an analysis and visualization tool. In the 3D synthetic image the differences between built reality and planned city are more apparent and easier to interpret.

Figure 8: Parcel with overbuilt and underbuilt fragments

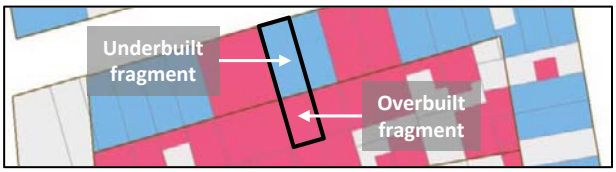


Figure 9: Overbuilt aggregated area in parcels

Figure 10: Underbuilt aggregated area in parcels

Figure 11: Height interpretation in the 3D model

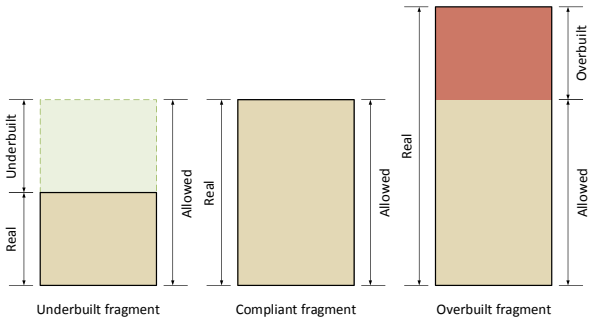
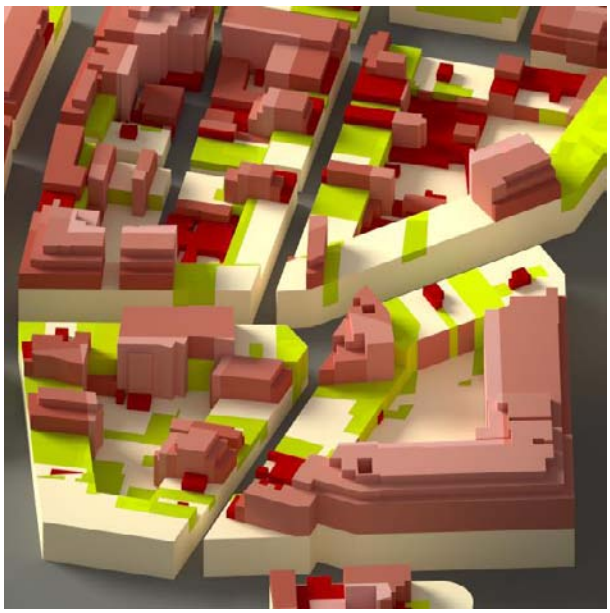


Figure 12: Volumetric analysis of overbuilt (red) and underbuilt (green) fragments in the case of study

Figure 13: Aerial axonometric view and 3D representation of the same area



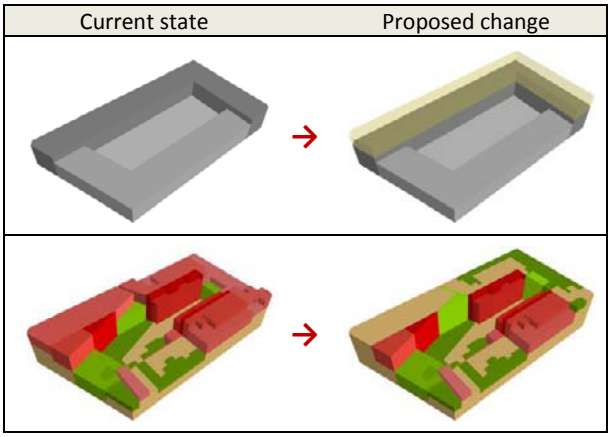
Source: Bing Maps

3.3 Quantification of the outcomes of proposed changes in regulations

In addition to the visualization of the morphological differences between the proposed city and the real built environment, the methodology allowed planners to precisely quantify the outcomes of proposed regulation changes.

Figure 14 shows an example of a proposed regulation change, where two additional floors could be allowed in part of a city block. The methodology allowed decision-makers to measure the additional built area that would become compliant with the new regulation and the number of properties that would change its legal status.

Figure 14: Proposed increased allowed height in a city block (above) and resulting outcome (below)



4 Conclusions

The presented methodological approach seeks to assist in implementing better policies in urban transformation processes, with better and easier to understand information, making possible to measure and visualize with precision the conformity of the built environment to the determinations of Urban Plan, and to evaluate the outcome of proposed regulation changes.

The 3D visualization techniques allow the discovery of patterns not obvious even for trained professionals and is a valuable tool to communicate the results of the analysis.

To improve the accuracy of the analysis, a methodology to study adjacency relations in a non-topological GIS was developed using SQL, which allowed assigning height values to entities that didn't have this attribute from their spatial context.

As further investigations, an improvement of the methodology is proposed to incorporate information about building quality, economic activity and demographic information in densification processes to be able to:

- Determine which parcels are more likely to be transformed according to their age, uses, habitability, economic value, etc.
- Estimate the potential number of people affected by and/or benefited from regulation changes.
- Calculate the taxation of the increased value of the properties in redevelopment scenarios.
- Prioritize zones with greater incompliance and/or obsolete typologies (such as outdated industries) to be included in transformation processes.
- Explain the possible historic reasons that have resulted in the current morphology of the city.

Furthermore, to improve the accuracy and usefulness of the visualization of the results it is proposed:

- The incorporation of a Digital Elevation Model (DEM) in the 3D model.
- The use of Augmented Reality tools to visualize the results on site.

Acknowledgements

The authors would like to thank the staff in the Urban Studies Bureau of the Urban Planning Department of the Barcelona City Council for their collaboration.

References

[1] A. Power, "Does demolition or refurbishment of old and inefficient homes help to increase our environmental, social and economic viability?," *Energy Policy*, vol. 36, no. 12, p. 4487–4501, 2008.

[2] The World Bank, World Development Report 2009: Reshaping Economic Geography, Washington DC: World Bank, 2008.

[3] A. Nelson, R. Burby, E. Feser, C. Dawkins, E. Malizia and R. Quercia, "Urban containment and central-city revitalization," *Journal of the American Planning Association*, vol. 70, no. 4, pp. 411-425, 2004.

[4] P. Garcia-Almirall y J. Roca Cladera, "A New Setting For Reflecting On Urban Matter. Integration Of GIS Technology In Advanced Internet," in *6th Agile Conference On Geographic Information Science*, Lyon, 2003.

[5] T. Yigitcanlar, "Australian Local Governments' Practice and Prospects with Online Planning," *URISA Journal*, vol. 18, no. 2, pp. 7-17, 2006.

[6] V. A. Ceccato and F. Snickars, "Adapting GIS technology to the needs of local planning," *Environment and Planning B: Planning and Design*, vol. 27, no. 6, pp. 923-937, 2006.

[7] S. Rueda Palenzuela, Modelos e Indicadores para ciudades más sostenibles, Barcelona: Departament de Medi Ambient de la Generalitat de Catalunya, 1999.

[8] J. Alguacil, A. Hernández, M. Medina y C. Moreno, La Ciudad de los ciudadanos, Madrid: Centro de Publicaciones, Ministerio de Fomento, 1997.

[9] Corporación Metropolitana de Barcelona, Normas urbanísticas: plan general metropolitano de ordenación urbana de la entidad municipal metropolitana de Barcelona, Barcelona: Corporació Metropolitana de Barcelona, 1976.

[10] M. Ribas Piera, Los denominados standards urbanísticos y su aplicación al planeamiento, Barcelona: Urbanística III, Escola Tècnica Superior d'Arquitectura de Barcelona, Universitat Politècnica de Catalunya, 1982.

Session:
Qualitative Information

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Qualitative Representation of Dynamic Attributes of Trajectories

Tales Paiva Nogueira Hervé Martin
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
CNRS, LIG, F-38000 Grenoble, France
paivan@imag.fr herve.martin@imag.fr

Abstract

Trajectory dynamic characteristics may be a very relevant source of information to analyze the behaviour of moving objects. However, most of existing works on trajectory representation deal only with basic parameters of trajectories, namely space and time. In this paper, we show how some derivatives of the spatio-temporal dimension, e.g. speed, acceleration, direction, may be integrated in trajectory modelling. We address the problem of representing trajectories in a way that qualitative descriptions of trajectories are stored and easily accessed through an ontology called QualiTraj which is also flexible enough to support relevant raw data representation. We validate our proposal with real GPS traces collected from a well-known sports tracking mobile application.

Keywords: geographical information systems, trajectory analysis, semantic trajectories, dynamic, user profile, ontologies

1 Introduction

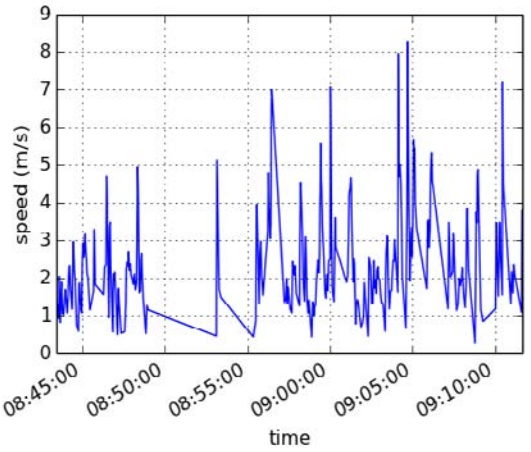
The Internet as we know it today is constantly evolving to a more and more connected system thanks to the increasing quantity of data made available as linked data, building what is called the Semantic Web [2]. With semantic web technologies we are moving from the web of documents towards the web of data, where machines will be able to understand and reason about the connections among different datasets and therefore enable the development of richer applications. It is of common knowledge that ontologies are well suited to represent datasets in this new paradigm.

At the same time, we are witnessing the development of mobile technologies such as smart phones for the acquisition of data in conjunction with many sensors as well as the grow of technologies of geolocation (GPS, A-GPS, GLONASS) and identification (RFID). The convergence of these technologies allows the easy acquisition of information about the trajectories of users using mobile devices. The acquisition, management, modelling, and analysis of such data provide many challenges related to the integration of these data with systems that already exist. Therefore, it should be taken into account that the multidimensional and multifaceted aspects of these data potentially holds a very rich semantics. There is a vast amount of works that propose to bridge the gap between trajectory representation and Semantic Web technologies, mainly regarding the representation with ontologies [1, 3, 9, 10, 13, 17, 19, 20].

The identification of mobility patterns has been an constant topic of interest in the GIScience area in several domains like tourism, road traffic, crisis management, marketing, etc. [8]. Several works have dealt with trajectory analysis proposing new ways of comparing, segmenting and clustering moving object's paths. But most of them only handle the geometric aspect of trajectories and just a few deal with dynamic parameters like speed and acceleration explicitly [4, 12].

In most cases, the variability of dynamic properties is very high. Take as example the raw representation of speed of a runner in Figure 1. Although it is obvious that the movement did not suffer exactly the same variations as it is depicted in the graphic due to the intrinsic error and noise of GPS readings, we can observe that there is a need to simplify this information so it can become more useful. The development of new methods and tools to analyse movement components like the one shown in Figure 1 is still a great challenge.

Figure 1: Time series of the speed profile of a runner captured by a smart phone application without any post processing treatment.



In this work, we argue that a qualitative representation of trajectories components are useful and enable new queries to be build and answer many application domain needs. The remainder of this paper is organized as follows: in section 2, we compare our work with similar proposals and highlight the differences between them. In section 3, we define what are the dynamic aspects considered in this paper. In section 4, the

QualiTraj ontology is introduced, followed by a case study in section 5.

2 Related work

In [14], Rehrl et. al. proposed a method for semantic processing of GPS traces where information is extracted from raw data. Based on the assumption that the basic parameters to express motion in space and time are velocity and course, they defined six motion patterns with associated rules. The patterns are the following: *stand still* characterizes the absence of motion and is assumed when the velocity is less than 1 m/s; *steady motion* represents the periods when there is motion with constant velocity and is distinguished when velocity is greater than 1 m/s and acceleration lies between −0.3 m/s² and 0.3 m/s². *Positive acceleration* happens when the velocity increases and is greater than 1 m/s and acceleration is greater than 0.3 m/s²; *negative acceleration* is similar but acceleration should be less than −0.3 m/s². *Positive course change* is identified when there is a course change rate above 0.4 º/s, and *negative course change* is determined when this change is below −0.4 º/s.

While this categorization has as objective to improve the level of abstraction of motion data, the authors rely too heavily in thresholds to characterize speed, acceleration, and course changes. In an heterogeneous dataset, this approach does not seem adequate as these thresholds may vary depending on the mean of transportation. In our work, we preferred the usage of statistic measures whenever it was possible to avoid relying on thresholds that depends on the nature of the data being analyzed.

In [9], van Hage et. al. presented the Simple Event Model (SEM) and its application in the maritime domain. In their use case, events are automatically recognized from the Automatic Identification System (AIS) raw data and represented as SEM instances. From that, it was possible to characterize some types of ship behavior like *slowing down*, *speeding up*, and *anchored*. Three types of data were collected in the form of time series: location, speed, and course. Due to the large dimensions of the tracked ships and the fact that they do not accelerate nor change their courses quickly, their movement are very regular and much more easy to compress by a piecewise linear algorithm. In our paper, instead of AIS data with speed information already included, we have at first just GPS raw data from which we have to calculate the speed profile. Moreover, the nature of motion data is very different: runners instead of ships. Runners may have a much more irregular speed, acceleration, and changes in course direction when compared to ships. Besides, we take a qualitative approach towards the characterization of movement.

3 What is dynamic?

One of the most important features of spatio-temporal systems is the ability to trace the path that a moving object follows during some time. A trajectory can be defined as the user defined record of the evolution of the position (perceived as a point) of an object that is moving in space during a given time interval in order to achieve a given goal [15]. A research topic that is constantly studied in the trajectory analysis domain is related to the representation of these spatio-temporal paths. While the representation of trajectories with ontologies have already been subject of many studies, the dynamic aspects of trajectories are generally just mentioned as simple attributes or even not mentioned. Most works about trajectory analysis limits themselves to the geometric representations of trajectories as a static curve [5].

The dynamic properties that we talk about in this paper may have different names among the literature. Dodge et. al. [5], for instance, call them movement parameters and separate them in three groups: primitive parameters, primary derivatives, and secondary derivatives. Each group is further organized in spatial, temporal and spatio-temporal dimensions. The primitive parameters are the ones that has been the subject of most studies in GIS (position and time). The primary derivatives are distance, direction, spatial extent, duration, travel time, speed and velocity. The secondary derivatives are spatial distribution, change of direction, sinuosity, temporal distribution, change of duration, acceleration, and approaching rate. In this work, we are going to focus on the speed derivative, as we believe that this aspect of trajectories is crucial to the characterization of the behaviour of a moving object.

4 The QualiTraj Ontology

In this section, we present a modeling approach that enables the representation of trajectories' dynamic characteristics in a high abstraction level through an ontology. Figure 2 shows the basic structure of QualiTraj, the main contribution of this work. The top level element is the *Trajectory* entity, which represents a spatio-temporal path followed by a moving object. Each trajectory may have one or more profiles. Each *Profile* represents one dynamic aspect of a trajectory (e.g. speed, acceleration, direction).

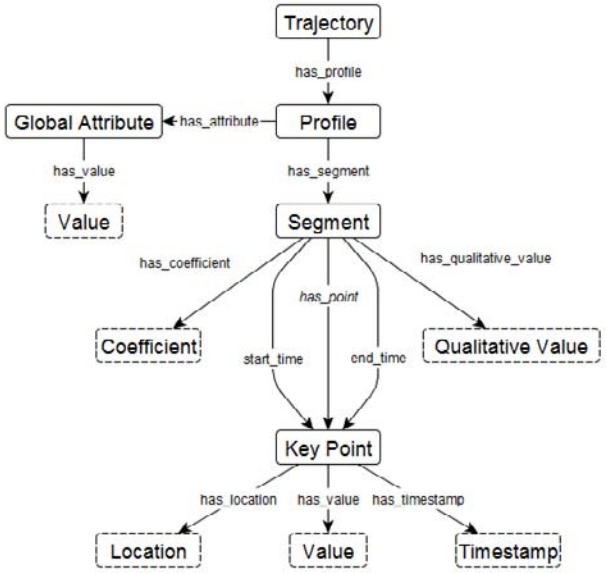
Profiles may have aggregated measures that might be useful depending on the application. Thus, we included the *Global Attributes* entity to store information like the average speed of the whole trajectory.

The *Segment* is the entity that represents a relevant change in the dynamic property occurred along the trajectory. This element contains the qualitative information itself stored in the *Qualitative Value* entity. Each *Segment* starts and ends at a *Key Point*, i.e. a location in space and time that define the bounds of the segment. The *Key Points* may also be used to retrieve important information, e.g. where and when the highest speed was achieved. While this kind of data is not mandatory because it is application-specific and, on the other hand, the start and end points must always be represented, there are three relationships between *Segment* and *Key Point* in the ontology being optional only the one called *has_point*.

The kind of change is stored in the *Qualitative Value* entity associated with each *Segment*. The application developer should determine which values compose the lexical space of this entity. Another important element to represent each *Segment* is the *Coefficient*, an entity that holds the slope of the line that connects the starting and ending points. Having this information may be useful if we want to infer the approximate

value of the profiled characteristic using a linear equation. The next section shows an example of the usage of the QualiTraj ontology in a real scenario.

Figure 2: The QualiTraj ontology



5 Case study

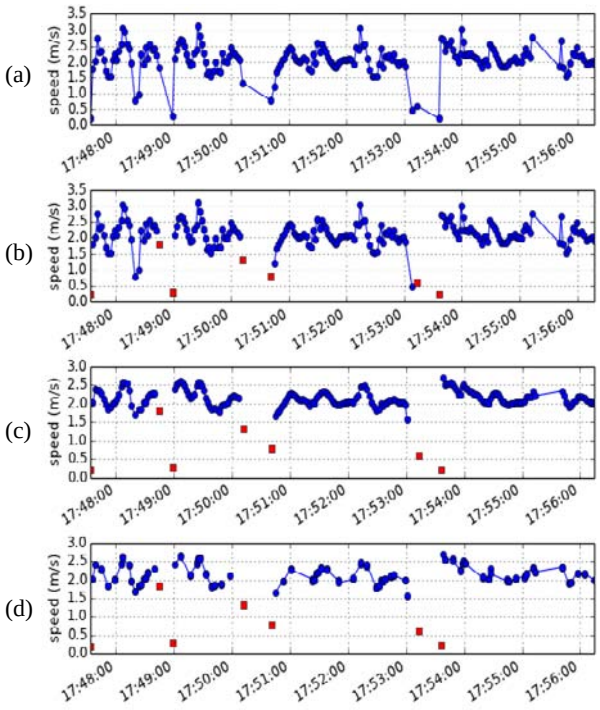
The studies about mobility analysis are numerous in the literature. In order to validate them, it is of vital importance to work with a representative dataset. The capture of real-life spatio-temporal data is generally expensive and time consuming due to the need of adequate equipment, search for subjects willing to participate (e.g. taxi drivers, shoppers, students), among other factors. Fortunately, there are some available datasets that can be freely downloaded, like GeoLife [21, 22], Reality Mining [6], geo-tagged photos from websites like Flickr[1] and Instagram[2], among others.

One interesting source of trajectory data is sports tracking websites and mobile applications, e.g. RunKeeper[3], Endomondo[4], Sports Tracker[5], Strava[6], MapMyRun[7], and have been source of studies like the one by Ferrari and Mamei [7]. Notwithstanding the widespread adoption of these services by professional athletes as well as by casual practitioners of sports activities, the collected data is not always publicly available. The minor part of sites provide an open API for third-party applications to access user-generated data.

For our case study, we collected data from users of MapMyRun application that shared their workouts publicly. We gathered information about 10 users that logged activities in the city of Grenoble, France, represented by 66 trajectories in total. In order to be clearer, we are going to show the raw data transformation steps of one short workout as it becomes easier to spot the changes suffered by the time series through all the steps. Figure 3 shows the raw speed data being pre-processed in order to simplify the stored data. The first graph shows the calculated speed at each point of the trajectory based on the latitude, longitude, and time difference between points. We have used the Haversine formula to calculate the approximate distance traveled during each sampled point. It is important to notice that a good speed approximation is heavily dependent on a good sampling rate, i.e. GPS fixes constantly recorded in small intervals of time.

The second step of the cleaning phase consists in detecting stops and moves of the tracked object. After that, we applied a Kalman filter [18] in order to smooth the data and attenuate GPS position errors. The last step of the smoothing phase is to summary the data points with a piecewise linear segmentation [11]. All the steps of the cleaning phase are depicted in Figure 3. In this specific example, the length of the time series was reduced from 60 points to only 28 points without losing the main characteristics of the signal.

Figure 3: The evolution of speed during a four-minute walk and the steps of post-processing: (a) is the raw data, in (b) stops and moves are identified, (c) is the filtered signal, and (d) shows the piecewise linear segmentation result.

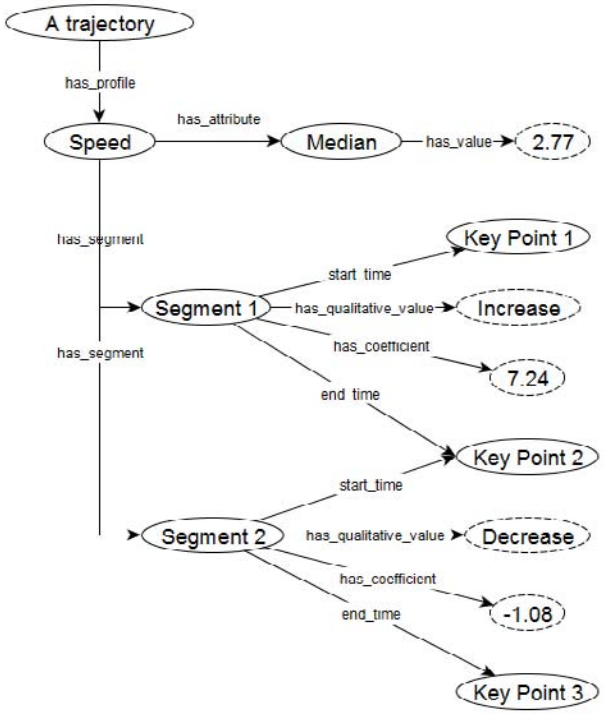


The final step in the speed representation of this trajectory consisted in creating the entities following the QualiTraj model. The lexical space used for the *Qualitative Value* entity was {"Increase", "Decrease", "Steady", and "Stop"}. Figure 4 shows the first two Segments represented with QualiTraj. The first segment consists in an increase of speed from 2.02 m/s to 2.40 m/s, which are the points of a line with an angle of 7.24 degrees. We omitted the timestamp and location of *Key* Points to improve the readability of the example. Notice that the

[1] www.flickr.com
[2] www.instagram.com
[3] www.runkeeper.com
[4] www.endomondo.com
[5] www.sports-tracker.com
[6] www.strava.com
[7] www.mapmyrun.com

same *Key Point*, "Key Point 2", has been reused in both segments, avoiding data duplication thanks to the graph structure of the ontological modeling approach.

Figure 4: Part of qualitative representation of a trajectory using the QualiTraj ontology



6 Conclusion

In this paper, we demonstrated how it is possible to enrich raw trajectory data with dynamic aspects of movement and provide an infrastructure for querying this new knowledge through an ontology.

The representation of spatio-temporal data by means of ontologies is even more useful when the inference features of reasoners are explored. As a following activity of this work, we will investigate how reasoners can improve the analysis of patterns of dynamic movement parameters of trajectories. Queries that involve more than one moving object form and important group of queries about relative motion and should be studied in the future.

Another important development will be the connection of the proposed ontology with different linked data sources like the LinkedGeoData project [16], which provides OpenStreetMaps information in the format suitable for the Semantic Web. For instance, we could formulate queries in the domain of traffic analysis to find drivers that do not slow down near traffic-calming features as the OpenStreetMaps dataset has a traffic-calming key for many possible features of this type, like bumpers, chicanes, cushions, and others. In this way, more complex queries and reasoning tasks can be also envisaged as future work.

Acknowledgments

The authors would like to thank the French Ministry of Higher Education and Research (Ministère de l'Enseignement Supérieur et de la Recherche de la France – MESR) for supporting this work.

References

[1] Baglioni, M., Macedo, J., Renso, C. and Wachowicz, M. 2008. An ontology-based approach for the semantic modelling and reasoning on trajectories. *Advances in Conceptual Modeling – Challenges and Opportunities*. Springer Berlin Heidelberg. 344–353.

[2] Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*. 5, 3 (Jan. 2009), 1–22.

[3] Camossi, E., Villa, P. and Mazzola, L. 2013. Semantic-based Anomalous Pattern Discovery in Moving Object Trajectories. *CoRR*. abs/1305.1, (May 2013).

[4] Dodge, S., Weibel, R. and Forootan, E. 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*. 33, 6 (Nov. 2009), 419–434.

[5] Dodge, S., Weibel, R. and Lautenschütz, A.-K. 2008. Towards a Taxonomy of Movement Patterns. (2008), 1–12.

[6] Eagle, N. and (Sandy) Pentland, A. 2005. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*. 10, 4 (Nov. 2005), 255–268.

[7] Ferrari, L. and Mamei, M. 2013. Identifying and understanding urban sport areas using Nokia Sports Tracker. *Pervasive and Mobile Computing*. 9, 5 (Oct. 2013), 616–628.

[8] Güting, R.H., Böhlen, M.H., Erwig, M., Jensen, C.S., Lorentzos, N.A., Schneider, M. and Vazirgiannis, M. 2000. A foundation for representing and querying moving objects. *ACM Transactions on Database Systems*. 25, 1 (Mar. 2000), 1–42.

[9] Van Hage, W.R., Malaisé, V., de Vries, G., Schreiber, G. and van Someren, M. 2009. Combining ship trajectories and semantics with the simple event model (SEM). *Proceedings of the 1st ACM international workshop on Events in multimedia - EiMM '09* (New York, New York, USA, 2009), 73.

[10] Hu, Y., Janowicz, K., Carral, D., Scheider, S., Kuhn, W., Berg-Cross, G., Hitzler, P., Dean, M. and Kolas, D. 2013. A Geo-ontology Design Pattern for Semantic Trajectories. *Spatial Information Theory*. T. Tenbrink,

J. Stell, A. Galton, and Z. Wood, eds. Springer International Publishing. 438–456.

[11] Keogh, E., Chu, S., Hart, D. and Pazzani, M. 2001. An online algorithm for segmenting time series. *Proceedings 2001 IEEE International Conference on Data Mining* (2001), 289–296.

[12] Laube, P., Dennis, T., Forer, P. and Walker, M. 2007. Movement beyond the snapshot - dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*. 31, 5 (2007), 481–501.

[13] Parent, C., Pelekis, N., Theodoridis, Y., Yan, Z., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A. and Macedo, J. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys*. 45, 4 (Aug. 2013), 1–32.

[14] Rehrl, K., Leitinger, S., Krampe, S. and Stumptner, R. 2010. An Approach to Semantic Processing of GPS Traces. *Proceedings of the 1st Workshop on Movement Pattern Analysis* (Zurich, Switzerland, 2010), 136–142.

[15] Spaccapietra, S., Parent, C., Damiani, M.L., Macedo, J.A.F., Porto, F., Vangenot, C. and Demacedo, J. 2008. A conceptual view on trajectories. *Data & Knowledge Engineering*. 65, 1 (Apr. 2008), 126–146.

[16] Stadler, C., Lehmann, J., Höffner, K. and Auer, S. 2012. LinkedGeoData: A Core for a Web of Spatial Open Data. *Semantic Web journal*. 3, 4 (2012), 333–354.

[17] Wannous, R., Malki, J., Bouju, A. and Vincent, C. 2013. Modelling Mobile Object Activities Based on Trajectory Ontology Rules Considering Spatial Relationship Rules. *Modeling Approaches and Algorithms for Advanced Computer Applications*. A. Amine, A.M. Otmane, and L. Bellatreche, eds. Springer International Publishing. 249–258.

[18] Welch, G. and Bishop, G. 1995. *An introduction to the Kalman filter*.

[19] Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S. and Aberer, K. 2011. SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories. *Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11* (New York, New York, USA, 2011), 259.

[20] Yan, Z., Macedo, J., Parent, C. and Spaccapietra, S. 2008. Trajectory Ontologies and Queries. *Transactions in GIS*. 12, (Dec. 2008), 75–91.

[21] Zheng, Y., Xie, X. and Ma, W.-Y. 2010. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin*. 49 (2010), 1–8.

[22] Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y. 2009. Mining interesting locations and travel sequences from GPS trajectories. *Proceedings of the 18th international conference on World wide web WWW 09*. 19, (2009), 791.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

VGI Edit History Reveals Data Trustworthiness and User Reputation

Fausto D'Antonio
University of L'Aquila
Via G. Gronchi 18, 67100
L'Aquila, Italy
fausto.dantonio@gmail.com

Paolo Fogliaroni
Vienna University of Technology
Gusshausstr. 27-29, 1040
Vienna, Austria
paolo@geoinfo.tuwien.ac.at

Tomi Kauppinen
Aalto University School of Science
Department of Media Technology
FI-00076 Aalto, Finland
tomi.kauppinen@aalto.fi

Abstract

Volunteered Geographic Information (VGI) is an approach to crowdsource information about geospatial features around us. People around the world are engaged with typing in their observations about the world (like locations of shops, cafeterias), or to semi-automatically gather them with mobile devices (like hiking paths or roads). In this process people might make mistakes, for instance assign misleading tags to features or provide over simplistic boundaries for features. In this paper we study what kinds of things might contribute to assess trustworthiness of data, and reputation of contributors for VGI. We present a model for analysing the different factors, and a method for automatically creating the trust and reputation scores.

1 Introduction

Volunteered Geographic Information (VGI) [5] is an approach to crowdsource information about geospatial features around us. Recently, VGI is gaining increasing attention and (web) services relying on it are becoming ubiquitous.

Thanks to the greater engagement of contributors, coverage and precision of VGI data is quickly approaching the level granted within professional Geographic Information Systems (GIS), as shown by several comparative studies with official national datasets [6, 14]. However, while in professional GIS data quality is granted by certified authorities, the assessment of VGI data quality remains an open challenge [11, 7, 13].

A basic method to assess the quality of a VGI dataset consists in comparing it against a professionally–generated ground–truth dataset. However, this approach suffers major drawbacks. First, it requires access to professional datasets that, in the best case, is expensive and, in the worst case, is not possible at all. Moreover, it does not provide a quality assessment procedure that is universally valid (e.g., think of cases where a ground–truth dataset is not available at all).

As suggested in [1], a different approach consists in assessing the quality of VGI data through a proxy measure: trustworthiness. Trustworthiness is defined [12] as a "bet about the future contingent action of others". In this sense, trustworthiness is strictly related to the concept of (others') reputation.

This paper presents ongoing work on an evaluation model of volunteers' reputation and data trustworthiness that derives the coveted information from VGI data, without requiring a comparison with external sources. We draw inspiration from the work in [7] and extend it by (i) relating data trustworthiness and user reputation and (ii) accounting for the relevance of data editing. (iii) Finally, our model accounts for atomic editing operations, rather than for composite editing patterns.

2 Related work

Quality assessment of VGI is still rather new research topic. As suggested by Flanagin and Metzger [3], there is a critical need for identifying methods and techniques to evaluate the VGI quality.

One rather standard approach is to compare VGI datasets to authoritative, ground-truth datasets, as done, for example, by Mooney, Corcoran and Winstanley [11] who analysed characteristics of polygons contributed by OpenStreetMap users. According to the results, volunteers seem to be able to more easily trace outlines of water features compared to forest features.

A different approach is undertaken by Bishr and Janowicz [1] that promote the notion of informational trust to be used as a proxy measure for quality. Their proposal was one of the first examples for using trustworthiness for quality assessment in VGI.

Keßler, Trame and Kauppinen apply the Bishr and Janowicz proposal to use trust as a proxy measure, in [8] they used trust and provenance for studying contribution patterns in the case of OSM. An extension of this work is [7] in which, Keßler and De Groot, provided a few indicators that influence trust and that were basically derived from data provenance.

The work presented in [7] has the purpose to build a model that depends mostly on provenance data; so that there is no need of a reference comparison dataset; trustworthiness is associated to each feature and represents the proxy value of data quality. In this work Keßler introduce the user reputation issue and leave it for future refinement.

Keßler used five parameters for trustworthiness evaluation. (1) Versions, they are an important source of provenance information. (2) Users, the higher is the number of users that works on a feature the higher is the trustworthiness value. (3)

Confirmations, all the revisions that were made in the neighbourhood of a feature are taken into account. (4) Tag corrections, a semantic change over a feature decreases the

feature trustworthiness. (5) Rollbacks, restoring a feature's previous state also decreases the feature trustworthiness.

3 Model Overview

In this section we give an overview of the main constituents and mechanisms underlying our trustworthiness evaluation model. More details are given in next sections.

The model we introduce can work with any VGI system that provides the following, basic requirements:

- The system supports (directly or indirectly) feature versioning
- which is expressible as a sequence of basic editing operations of the type: creation, modification, and deletion.

3.1 Feature Versioning

Geographic features in a VGI system are subject to repeated changes over time. A change is operated by a contributor and brings a feature representation into a new state or version. History of a feature's changes is referred to as feature provenance [8] and feature versioning is a particular interpretation of it.

Similarly to the work presented in [7] we base trustworthiness evaluation on provenance information. This approach is a realization of the "many eyes principle" [6] that assumes that incorrect, wrong, or malicious information about a feature get corrected over versions contributed by different volunteers. The main underlying idea is that a high number of lay contributors reporting on the same feature and an iterative adjustment of the feature information provides a valid alternative to field expertise.

Our approach consists in assigning a trustworthiness value to each version of a feature. Thus, in order to be compliant with our model, a VGI system must provide provenance information directly in the form of feature versioning. Alternatively, versioning must be derivable from feature provenance.

3.2 Editing types

While the approach presented in [7] derives trustworthiness values from editing patterns[1], our approach relies upon atomic editing types:

Creation When a new feature is inserted into the dataset there are no previous versions that the feature can be compared with. Thus, it is assigned a trustworthiness value equal to its author reputation (cf. Section 5.1).

Modification The modification of an existing feature yields a new version whose trustworthiness depends on the author reputation and on the compatibility with previous versions. That is, if previous versions are similar to current version the latter is associated a high trustworthiness. Contrarily, dissimilarities are rated with low scores. Note that a modification also affects trustworthiness of previous versions.

Deletion A deletion occurs when a real–world feature somehow disappears (e.g., a building is demolished). Notably, this editing type yields a new "void" version. Trustworthiness of this version is set equal to its author reputation and previous versions are not affected.

Studying how our approach behaves with respect to editing patterns is left for future work. However we would like to anticipate one notable situation. A deletion and a creation happening consecutively must be deemed a modification: this is the case that a feature is deleted and recreated rather than being modified. This also covers the case of non–genuine deletions.

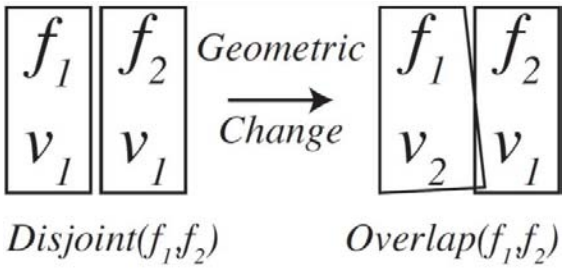
4 Measuring Editing Relevance

We argue that diverse edits must be weighted differently. More specifically, the more closely the version resulting from a change fits the feature in the real world, the higher the trustworthiness of this version.

Since a direct comparison with real world cannot be performed, we suggest evaluating the level of fitness by clustering the versions of a feature according to three main characteristics and comparing versions assigned to the same cluster.

A spatial feature consists of two components: spatial and semantic. Moreover, the spatial characteristic can be further refined into qualitative and geometrical. The reason for such a finer distinction relies on the fact that a small change in the geometry of a feature may correspond to a notable change in the qualitative spatial relations holding among this feature and its neighbours.

Figure 1: Geometric and qualitative spatial change.

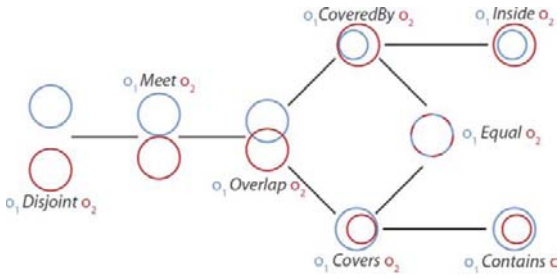


For example, let us consider the scenario depicted in Figure 1 where two features f_1 and f_2 at version v_1 are *disjoint* but very near to each other. A geometric change occurs that slightly modifies the geometry of f_1 as depicted on the right side of the figure. This small geometric change modified the topological relation holding among the two feature into overlap. Qualitatively, this is a notable change since according to the theory of *conceptual neighbourhoods* [4] the two relations are not close. Indeed, as shown in Figure 2: Conceptual neighbour graph of 9-Intersection model [2]., one cannot switch from the *disjoint* relation to the *overlap* relation

[1] An editing pattern is a sequence of atomic editings that can be interpreted as a unique high–level change. For example, it has been shown [8] that emerging editing patterns in OpenStreetMap (http://www.openstreetmap.org/) are confirmations, corrections, and rollbacks.

without going through the relation *meet*. Conversely, big geometric changes may not alter the qualitative arrangement of features. Thus, the consideration of both the geometric and qualitative aspect allows for mitigating the evaluation of spatial changes.

Figure 2: Conceptual neighbour graph of 9-Intersection model [2].



Accordingly, we distinguish the following characteristics:

Semantic The semantic or thematic aspect describes, by means of textual tags, the function of a feature in the world. When a semantic change occurs we evaluate its relevance by considering the semantic distance between the tag associated to the new version and the tags associated to versions in the same cluster. This can be done, for example, by considering the shortest path on a wordnet [10] graph[2]. Alternatively, the shorter conceptual distance in an ontology including both previous and altered tag can be used.

Geometric The geometric aspect describes the shape and position of the feature in the world. The relevance of geometric changes is evaluated with respect to a series of quantitative variances like, area, perimeter, and vertices number and position.

Qualitative (Spatial) This aspect addresses qualitative spatial relations of different types (e.g., topological, directional, distance) holding among a feature and its neighbours. The relevance of the change depends on the distance on the conceptual neighbourhood graph between the relations holding for the new version of the feature with respect to those occurring on previous versions.

5 Reputation and Trustworthiness

We denote trustworthiness and reputation by T and R, respectively. As done in [9], we bound the values of the two parameters between 0 and 1: $0 \leq T; R \leq 1$. We associate trustworthiness to each version v of a feature f: by $T(f_v)$, we denote the trustworthiness value of version v of feature f.
The reputation of a user u changes over time t: by $R(u,t)$, we denote the reputation of user u at time t.

[2] http://graphwords.com/

5.1 Reputation

User reputation depends on the trustworthiness of all the feature version his editing produced and is defined as the average of such values:

$$R(u,t) = \frac{\sum_{f_i \in F(u,t)} T(f_i)}{|F(u,t)|} \quad (1)$$

where $F(u,t)$ is the set of all the feature versions edited by user u until time t.

5.2 Trustworthiness

The overall trustworthiness value T of a feature version f_v accounts for three different effects: direct T_{dir}, indirect T_{ind}, and temporal T_{time}.

Direct Effect The parameter T_{dir} is the expression of the level of similarity with respect to the characteristics discussed in Section 4 between the current feature version f_v and previous ones. Accordingly, its value depends on three factors: semantic $T_{dir,s}$, geometric $T_{dir,g}$, and qualitative $T_{dir,q}$.
Direct effect is modelled as:

$$T_{dir} = w_s \cdot T_{dir,s} + w_g \cdot T_{dir,g} + w_q \cdot T_{dir,q} \quad (2)$$

where w_s, w_g, and w_q are weights used to balance the influence of the three characteristics.
To assure $0 \leq Tdir \leq 1$ we enforce:

$$w_s + w_g + w_q = 1 \quad (3)$$

and

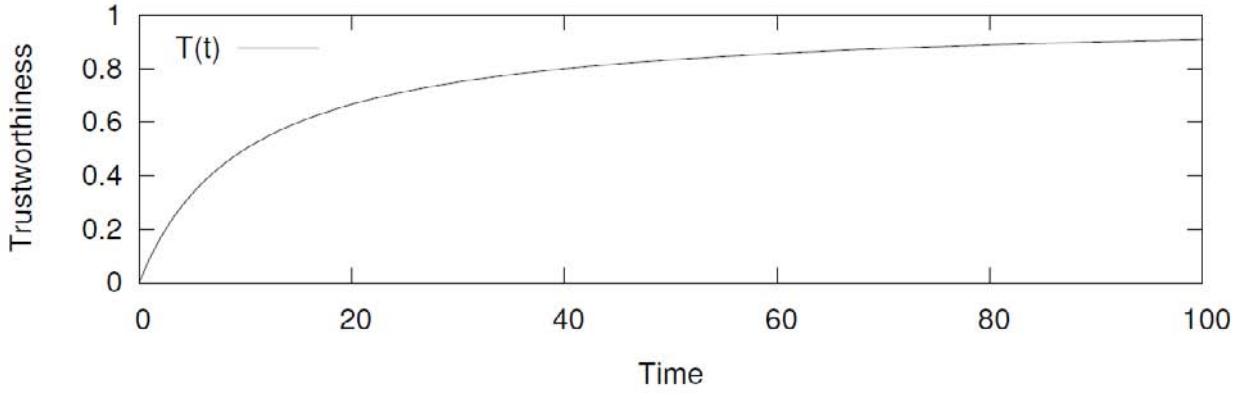
$$0 \leq T_{dir,s}, T_{dir,g}, T_{dir,q} \leq 1 \quad (4)$$

Indirect Effect The parameter T_{ind} models contributions on the overall trustworthiness value T that do not directly depend on the feature version f_v itself. For example, this can be used to account for confirmations [8]: the fact that a user contributes information about f_v's neighbours can be interpreted as a confirmation that f_v has a high fitness level with respect to the feature in the real world. Hence, f_v's trustworthiness must be increased.
Also in this case we account for three different factors, expression of the characteristics defined in Section 4:

$$T_{ind} = w_s \cdot T_{ind,s} + w_g \cdot T_{ind,g} + w_q \cdot T_{ind,q} \quad (5)$$

Similarly to direct effect, we assure T_{ind} falls in the interval [0; 1] by enforcing conditions similar to those reported in Equations 3 and 4.

Figure 3: Time effect on trustworthiness.



Temporal Effect The parameter T_{time} accounts for the effect of time on features trustworthiness. Namely, the longer a feature version f_v persists over time, the higher the probability that f_v has a high fitness level with respect to the real feature. Thus, if a feature version remains unaltered over time, its trustworthiness must be increased. We allow this by modelling time effect as:

$$T_{time} = \frac{t_v}{t_f + c} \tag{6}$$

where t_f is the life time of feature f (all versions), t_v is the life time of version v of feature f, and c is a parameter taking positive values that can be used to adjust the slope of the resulting curve (cf. Figure 3: Time effect on trustworthiness.). So, when t_v approaches infinity also t_f does, c becomes negligible, and T_{time} approaches 1.

Accordingly, the overall trustworthiness T is defined as:

$$T = w_{dir} \cdot T_{dir} + w_{ind} \cdot T_{ind} + w_{time} \cdot T_{time} \tag{7}$$

where w_{dir} and w_{ind} are weights used to balance the importance of the direct and indirect effect, respectively, and such that:

$$w_{dir} + w_{ind} = 1 \tag{8}$$

Temporal effect is weighted by

$$w_{time} = 1 - (w_{dir} \cdot T_{dir} + w_{ind} \cdot T_{ind}) \tag{9}$$

in order to assure that the overall trustworthiness value T increases as time passes with a pace following the curve depicted in Figure 3: Time effect on trustworthiness..

6 Conclusions and Outlook

Inspired by the work in [7], we introduced a model to evaluate VGI user reputation and VGI feature trustworthiness as a proxy measure for data quality.

We anchored our model to basic editing types and discussed how changes among feature versions can be evaluated grounding upon three characteristics: semantic, geometric, and qualitative.

We provided a high–level formulation for reputation and trustworthiness and discussed how the latter is a product of direct and indirect effects as well as a function of time.

The model is still under development, yet finer detailed than what was possible to discuss in this short paper. In the next phases we plan to implement the model and to study its behaviour using OpenStreetMap historical data. Also, a comparison with other trustworthiness evaluation models will be carried out.

References

[1] M. Bishr and K. Janowicz, Can we trust information? The case of volunteered geographic information, in *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume 640*, 2010.

[2] M. J. Egenhofer, A formal definition of binary topological relationships, in *Foundations of data organization and algorithms*, Springer, 1989, pp. 457-472.

[3] A. J. Flanagin and M. J. Metzger, The credibility of volunteered geographic information. *GeoJournal*, 72.3-4, pp.137-148, 2008.

[4] C. Freksa, Conceptual neighborhood and its role in temporal and spatial reasoning, in M. Singh e L. Travé-Massuyès, editors, *Decision Support Systems and Qualitative Reasoning*, pp. 181-187. North Holland, Amsterdam, 1991.

[5] M. F. Goodchild, Citizens as sensors: the world of

volunteered geography, *GeoJournal*, vol. 69, n. 4, pp. 211-221, 2007.

[6] M. Haklay, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets, *Environment and planning. B, Planning & design*, vol. 37, n. 4, p. 682, 2010.

[7] C. Keßler, J. Trame, and T. Kauppinen, Tracking editing processes in volunteered geographic information: The case of OpenStreetMap, in *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, 2011.

[8] C. Keßler and R. T. A. de Groot, Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap, in *Geographic Information Science at the Heart of Europe*, Springer, 2013, pp. 21-37.

[9] N. Mezzetti, A socially inspired reputation model, in *Public Key Infrastructure*, Springer, 2004, pp. 191-204.

[10] G. A. Miller, WordNet: a lexical database for English, *Communications of the ACM*, vol. 38, n. 11, pp. 39-41, 1995.

[11] P. Mooney, P. Corcoran, and A. C. Winstanley, Towards quality metrics for openstreetmap, in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[12] P. Sztompka, Trust: A sociological theory, Cambridge University Press, 1999.

[13] D. Zielstra, H. H. Hochmair, and P. Neis, Assessing the Effect of Data Imports on the Completeness of OpenStreetMap - A United States Case Study, *Transactions in GIS*, vol. 17, n. 3, pp. 315-334, 2013.

[14] D. Zielstra and A. Zipf, A comparative study of proprietary geodata and volunteered geographic information for Germany, in *13th AGILE international conference on geographic information science*, Leuven, 2010.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014, ISBN: 978-90-816960-4-3

A flexible framework for assessing the quality of crowdsourced data

Sam Meek
University of Nottingham
Triumph Road
Nottingham
sam.meek@nottingham.ac.uk

Mike J Jackson
University of Nottingham
Triumph Road
Nottingham
mike.jackson@nottingham.ac.uk

Didier G Leibovici
University of Nottingham
Triumph Road
Nottingham
Didier.leibovici@nottingham.ac.uk

Abstract

Crowdsourcing as a means of data collection has produced previously unavailable data assets and enriched existing ones, but its quality can be highly variable. This presents several challenges to potential end users that are concerned with the validation and quality assurance of the data collected. Being able to quantify the uncertainty, define and measure the different quality elements associated with crowdsourced data, and introduce means for dynamically assessing and improving it is the focus of this paper. We argue that the required quality assurance and quality control is dependent on the studied domain, the style of crowdsourcing and the goals of the study. We describe a framework for qualifying geolocated data collected from non-authoritative sources that enables assessment for specific case studies by creating a workflow supported by an ontological description of a range of choices. The top levels of this ontology describe seven *pillars* of quality checks and assessments that present a range of techniques to qualify, improve or reject data. Our generic operational framework allows for extension of this ontology to specific applied domains. This will facilitate quality assurance in real-time or for post-processing to validate data and produce quality metadata. It enables a system that dynamically optimises the usability value of the data captured. A case study illustrates this framework.

Keywords: crowdsourcing; data quality; quality assurance; quality control; location based services; dynamic surveying

1 Introduction

The concept of citizens as sensors is becoming broadly utilised as collection-enabling technologies are widely adopted in consumer devices. As a consequence, the term *crowdsourcing* is generic, and describes an array of different activities carried out by people in an *active* (*e.g.* filling out a survey) or *passive* (*e.g.* information mined from Twitter) sense.

Types of crowdsourcing range from highly organized methods of harnessing the collective power of the crowd, for example Amazon's Mechanical Turk (Kittur, et al. 2008) and other monetary reward based schemes (Horton and Chilton, 2010), to volunteered geographic information (VGI) such as OpenStreetMap (Haklay and Weber, 2008).

Citizen science (Aoki et al. 2008) is also a form of crowdsourcing that has an established history. It often requires an in depth knowledge of a project, and so can be considered a specialised case of crowdsourcing.

Data collected by volunteers is no longer confined to the desktop as mobile technology and smartphone capabilities allow for real-time acquisition of geolocated data. Mobiles also enable real-time sharing of the information and analysis of the data captured. These location-based tasking activities have been extensively utilised in ecology, e.g., iSpot[1], which uses participant experts and ratings system to identify wildlife through location-tagged photography. The use of passive crowdsourcing in location-based tasks has been seen in monitoring traffic flow in Google Maps[2] where a device running the software sends back anonymised data to a centralised repository. This is an example of a *producer model* set of quality elements as described by GeoViQua (Yang et al. 2012), defined in ISO19157 (ISO 2002). The *user/consumer model* is introduced in Diaz et al. (2012) corresponding to feedback reports and measures, which describes quality information for an existing dataset sourced from the crowd.

The focus of this paper is to present a framework for validating and assessing the quality of data contributed by citizens with a geographic component. Proactive data improvement through stimulation of authoritative data and metadata is utilised increase accuracy and reduce uncertainty. The standards described for data quality (ISO 19157) and for geospatial metadata (ISO 19115) (together with additional GeoViQua elements) are relevant as the stakeholder overseeing crowdsourcing activities acts as a data producer, but does not fully control the data measurement process. Additionally the stakeholder is able to make judgements and evaluate the data from their own perspective and can also harness dynamic interaction with the user to influence the way the data are captured. Therefore, additional quality elements incorporating a *stakeholder model* are needed to fully qualify the collected data. These elements derive from assessment concerning the user, like sensor accuracy linked to calibration measures, data captured in relation to other knowledge (Pawlowicz et al. 2011), or their interaction seen as sources of uncertainty (Rousell et al. 2014).

[1] www.ispot.org.uk

[2] maps.google.com

Our Quality Assurance (QA) framework allows for the derivation of three types of metadata corresponding to the three models through Quality Control (QC) checks, tests or measures. We explore this model through a case study on citizen observations of flooding (see COBWEB flooding case study[3]).

2 Quality of crowdsourced information

Data collected by the crowd often lacks metadata about its quality that can lead to it being disregarded by scientists (Alabri et al. 2010), however it can frequently complement or update authoritative surveys (Jackson et al. 2010). A prevalent issue within crowdsourcing is the ability to verify and validate data collected by participants, directly contributing to the assessment of the data quality of some existing authoritative dataset (Foody and Boyd 2012). At the same time, authoritative data can be used to control the validity of volunteered information (Comber et al. 2013). An alternative to assess the quality of volunteered information is to employ experts as validators (See et al. 2013).

Several methods of gaining knowledge about the quality of citizen collected data have been proposed; they include using a majority decision or control group (Hirth et al. 2012), using a reputation system (Alabri et al. 2010), (Clow et al. 2011), and using user mobility patterns with their previous quality to assess credibility of the contributed data (Mashhadi and Capra, 2011). A different approach is to attempt conflation of the citizen collected data with an authoritative source, such as OpenStreetMap and Ordnance Survey GB (OSGB) Open Data (Pourabdollah et al. 2013).

Metadata about data quality plays an important role when attempting to conflate limited authoritative and crowd sourced data in regions that do not have resources to produce complete authoritative data, such as Iraq (Fairbairn et al. 2013). Analysing the ISO 19157 metadata standard, data quality can be split into two main categories: internal quality, which refers to aspects such as completeness, attribute accuracy, positional accuracy and consistency, and external quality such as fitness for use (Wang et al. 1996, Brown et al. 2012, Li et al. 2012).

2.1 Three quality models

The stakeholder model proposed in the introduction sits between internal and external quality as a source of uncertainty linked to the user and their device(s). If the QA/QC framework is aimed at producing metadata about spatial data quality in the form of the ISO 19157 (the producer quality model), this process requires other types of quality elements. Table 1 describes an overview of quality elements that are considered as part of the QA process, with a focus on *active volunteers*.

Table 1: Quality elements for the stakeholder quality model

Quality element	**Definition**
Vagueness	Inability to make a clear-cut choice (*i.e.*, lack of classifying capability)
Ambiguity	Incompatibility of the choices or descriptions made (*i.e.*, lack of understanding, of clarity)
Judgement	Accuracy of choice or decision in a relation to something known to be true (i.*e.*, perception capability and interpretation)
Reliability	Consistency in choices / decisions (*i.e.*, testing against itself)
Validity	Coherence with other people's choices (*i.e.*, against other knowledge))
Trust	Confidence accumulated over other criterion concerning data captured previously (linked to reliability, validity and reputability)

3 A generic quality assurance framework

A framework is required for quality assurance to understand and improve quality in crowdsourced data, with a view to increasing the quality of the entire database over time through directed data collection and error reduction. During this process, quality metadata values for the producer model, the consumer model and the stakeholder model are derived.

In a more general context, the stages for validation constituting the QA may be thought of as a series of discrete processes that could be flexibly (and iteratively) called under the control of a business process execution design that is specific to a case study but derived from generic principles. We have designed a QA process based on authoring a workflow for each type of data collected. The system is enabled by the Workflow Quality Control Authoring Tool (WoQC-AT) for chaining quality processes.

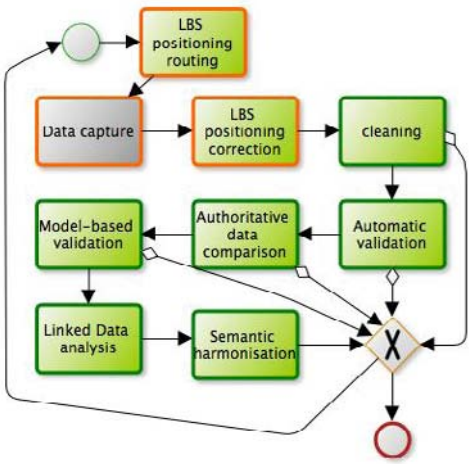


Figure 1: Typical workflow for quality assessment of crowd-sourced data before and after data capture (BPMN diagram)

An OGC compliant Web Processing Service (WPS) enables the execution of each QC element. It also composes a workflow using a back-end QA/QC service for the crowdsourcing data assessment.

[3] http://cobwebproject.eu/

The metadata for each process within the WPS is enriched by an *ontology* enabling retrieval of the appropriate processing checks, (WoQC-O). Figure 1 shows typical top-level stages for a stakeholder user-defined instance where each step encapsulates a sub-workflow. The top-level workflow includes a position quality improvement step before the data capture but only the green boxes are registered in the WoQC-WPS as the mobile app can also perform QA in certain circumstances.

Generally, the stages for validation and QA are discrete processes that can be flexibly (and iteratively) called under the control of a business process execution stage that may be either generic (by default), or use-case specific.

4 Ontology of quality assements

The QA/QC framework is built upon seven pillars of validation and quality assessment. These pillars cover aspects that can be a cause for concern with respect to quality when acquiring crowd data collected from mobile handheld devices in the environment.

This generic set of checks is chosen to illustrate the most suitable options available. Each of the sections encompasses a range of known techniques, some of which have previously been employed in crowd-sourcing projects and described in the literature. The purpose of the WoQC-O ontology (Figure 2) is to organise these techniques to perform iterative uncertainty reduction and accuracy improvement to facilitate authoring of the QA by instantiation of a workflow on a server.

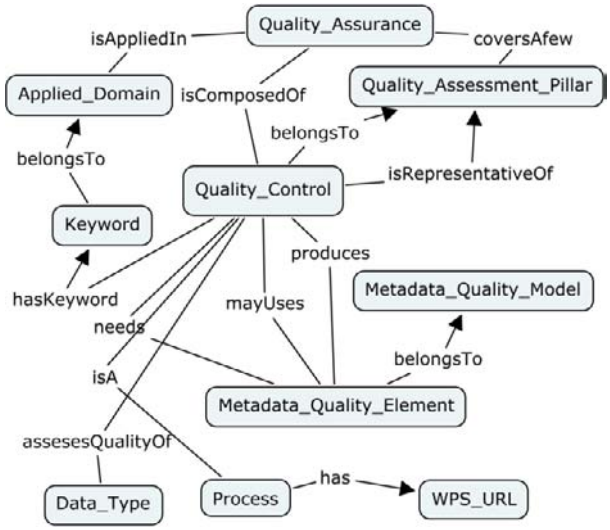


Figure 2: Top levels of the WoQC-O ontology (conceptual map diagram)

The following sub-sections detail the pillars in turn; each one combines a few checks or quality assessments that are processes registered in the WPS and seen as basic workflow. Figure 3 describes a generic QC single process with data inputs from authoritative sources (orange), crowdsourced inputs (green) and other inputs (grey) with their existing metadata.

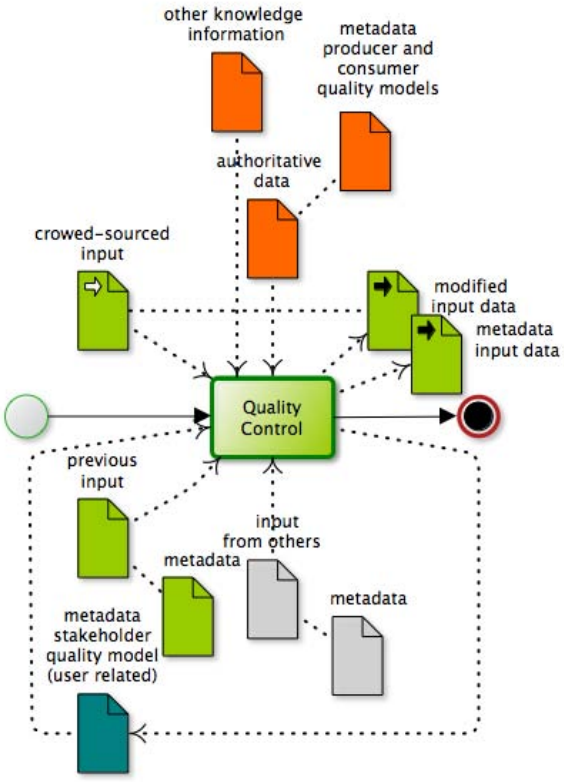


Figure 3: Generic atomic workflow QC process within the WoQC-WPS (BPMN diagram)

4.1 LBS positioning

Using LBS techniques such as geofencing (Martin et al. 2011) and remote logging and query via line of sight (Meek et al. 2013a), (Meek et al. 2013b), a mobile app is used to direct the user towards parts of a study area that are of interest to project organizers. Depending on the study, this can prevent data being captured when the positional accuracy is too low, it can also help to increase the density of observations where required, and can partially address the sampling problem in crowdsourcing.

From a quality perspective, this pillar is likely to minimise errors in recording field data as the user has few choices for data input. Additionally, asking a user to simply confirm or deny the existence of a potential observation requires little cognitive load on the part of the user.

4.2 Cleaning

Garbage removal and *data cleaning* uses low-cost checking mechanisms to remove erroneous entries, however there is a danger that valid data are discarded in this step. One level of garbage removal concerns false alarm data, or malicious entries. If crowdsourced data received has a capture position clearly outside of a study area, it can be removed immediately.

Besides rejection, data cleaning can also make the information collected more useful and suited to future stages outlined below. One such example is Stop Word Removal. Stop words are words that appear in text but have little meaning such as "and, &, a, the" (Barbier et al. 2012). Removal of such words is likely to help with stages applied later in the process such as such as conflation and semantic harmonization.

4.3 Automatic validation

In this stage, the data are assessed via automatic, computational techniques that apply a preliminary credibility check to the data collected. An example of employing these techniques is the OSMGB project where the aim was to check road names in OSM against the names released in the OS Open Data initiative as well as correcting the topology (Pourabdollah et al. 2013). The findings included the rate of error for OSM road labels is somewhat inversely proportional to the density of roads shown in the mapping. Validating topological relations between datasets, as a prerequisite for low level conflation has been a requirement in GIS technologies for sometime.

For an attribute manually input by the user, an *attribute range check* may relate to some obvious misunderstanding of units, as could automatic correction of spelling.

4.4 Authoritative data comparison

The purpose of this set of QC is to compare the collected data with authoritative data sources. This stage can be used to improve the confidence and validity of collected data, add attribution, and assign error bounds to the spatial, temporal and thematic attribute of a data item.

Research has focused on the user validating or updating authoritative data, e.g. (Foody et al. 2013) who describe a process where users add or change information on land cover data and Du et al. (2012), who use distributed logic to integrate crowdsourced vector road data with authoritative data.

The reverse view is to use authoritative data to validate the crowdsourced observations. Therefore the final validation process takes place after the quality assessment is done and a conflated dataset produced. Some of the quality elements for the crowdsourced data depend on other data sources, controlled by reference to a time stamp (*e.g.,* other crowdsourced data from Model-based validation). Records in the database enabling multiple representations are therefore tagged with a time and quality of real-world representation.

4.5 Model-based validation

This set of QC is focused on comparison of the crowd data with data from models or previously validated crowdsourced data. Models are likely to be environmental, but can also refer to different ways of prompting the users to harness contextual input. For environmental models it assesses the discrepancy between crowd inputs and model predictions.

Validation through directing the user geographically and through feedback of potential items of interest is assessed dynamically. The principle of improving quality by real-time data feeds and corrections (Pawlowicz et al. 2011) requires a server connection, a well-designed mobile application, or an ad hoc network between devices. Data collectors in the field are acting as a team without being aware of the other team members, and are in a sense multiple sensors, used to improve accuracy of a measurement.

The community of users, from the casual user to the domain expert can be used to derive a trust metric and personalise pushed tasks. Should the system know this information through a signup system, domain experts can be consulted to validate an observation if required.

4.6 Linked data analysis

Here, the term; linked data is being used in a broad sense and not just referring to Resource Description Framework (RDF) triplestores/databases. This stage combines the wealth of freely available data (big data) and associated data mining techniques to establish confidence and quality bounds for data inputs. Publicly available feeds such as Twitter are employed as a reference to newly captured information. Semantic accuracy plays a role and the coherence of the semantic information as defined in the stakeholder quality model (vagueness, ambiguity and validity elements) are used and also fed back into the metadata.

The different sections of QC can interact principally via the metadata, but also more complex workflows may involve a decision, validation and input of quality for a captured data element. This can be based on the conjunction of assessments from authoritative comparison and linked data analysis. For example, within a flooding event case study quantitative data captured may be assessed as poorly representative of the authoritative distribution, but Tweeted many times in the same time frame either in upstream or downstream of the location.

4.7 Semantic harmonization

This stage in the workflow illustrates methods of semantic integration of the crowdsourced and authoritative data. The set of QCs are transformations of the input data, ensuring conformance to or enrichment of an ontology, dependent on the application and domain.

A related method that can be used in preparation to harmonise to a specified ontology is through knowledge extraction and semantic similarities in VGI (Ballatore et al., 2013). In this two-stage process, the authors develop an OSM semantic network via a web-crawler and then produce a study where they look at the cognitive plausibility of different co-citation algorithms. This approach offers a system the ability to harmonise data entries with a crowdsourcing data repository (Idris et al. 2014).

5 Examples

The proposals presented above have been tested against a use case from the EU FP7 Project, COBWEB[3]. In this use case the citizen is asked to give some categorical and open textual information about the observation with instructions: flood height, speed and colour of stream as compared to three calibrated images of stream flows, free text and an image via the device's camera.

For simplicity, only one specific QC is mentioned here. Different quality checks may be used for different data types as highlighted by the shading in Table 2 but data may require the full set of checks to assess different scenarios.

Table 2: Flooding QA/QC scenario

	Activity	**Pillar check**/ specific QC	Outcome / **metadata**
1	User has reported a flood with details but no picture was taken.	**LBS positioning correction** / relative position of user and potential flood source	Geolocation of data captured with accuracy from the device, and flood source position with accuracy / **producer model**: spatial and temporal accuracy; thematic accuracy on the location name (place or river)
2		**Cleaning** / check data entry completion (content and position to the reported object)	The user is asked to get closer, if this is safe and to take a picture. / **producer model**: logical consistency **stakeholder model**: ambiguity, reliability, vagueness, judgement derived from the accuracy of the location name
3	The user gets closer and takes picture added to his/her previous record. (rechecking for LBS positioning and cleaning of step 1 and 2)	**Automatic validation** / image quality analysis: distance, resolution and focus optimising distance to take a picture	Estimated distance to flood source and optimum distance for photo report estimated are validating the record of sufficient quality / **producer model**: domain consistency, **stakeholder model**: trust
4	Data of judged flood high is checked against a DTM and flood model with historical data	**Authoritative data comparison** / Attribute data in the range of expected measures (within 2 standard errors of historical average)	Check for propensity for area to flood. Data value is borderline; a more real-time event validation needs to be performed for confirmation. / **producer model**: attribute accuracy takes the conflated variance, and sample of most spatiotemporal closed values; **stakeholder model:** validity, trust **consumer model**: (automatic) feedback report, rate of agreement
5	Other users that have recently contributed data from area are used for comparison and available (on-line) users are informed of for flood checking nearby.	**Model-based validation** / Attribute data in the range of recently observed data (within 2 standard deviations of the recently validated observations)	A similar trend is observed and the data captured is validated. / **producer model:** attribute accuracy takes the conflated variance, and sample of most spatio-temporal closed values; **stakeholder model**: validity
6	To increase credibility of the coverage accuracy of the flood over time, Twitter feeds are mined to check for recent reports of flooding	**Linked data analysis** /	A dataset of geolocated and temporally related tweets is created. Evidence of flood is computed by metrics such as #with_flood /#tweets, or other semantic measures. / **stakeholder model:** validity, judgement
7	The place name of this data point is checked against other recorded data points	**Semantic harmonisation** / similarity of names with known names and standard names	Standard place name and its variations such as local language and informal name is recorded at dataset level/ **producer model:** non-quantitative attribute sample of different values and similarity to the most commonly used **stakeholder model:** validity updated

At step 5, an estimate of the temporal distribution of the flood event may be inferred and this is controlled at step 6 where here only related information (not the flood height as in step 5) is compared.

All previous records from the same user may be used as well as its metadata to moderate the decisions made and also to modify the stakeholder metadata elements vagueness, ambiguity and reliability. At step 7, the collection of place names is useful for tweet mining for example.

6 Discussion and conclusion

The focus of the paper has been to present a framework in which QA/QC for assessing the credibility of crowd sourced data and enriching it to optimise user requirements can be facilitated. The required set of quality metadata has been identified and seven pillars in which the quality controls can occur have been described. Using the framework by authoring a workflow combining and chaining checks and quality assessments seen as processes belonging to the seven pillars provides the QA/QC for a crowdsourcing case study. The pillars represent the top levels of an ontology of quality controls that can be used. The ontology allows seamless access to appropriate QC when composing the workflow. Interoperability mechanisms of using standards such as WPS, BPMN, and the SKOS language to represent the ontology used to enrich the metadata of the WPS can ensure sharing of specific quality controls as processes.

Acknowledgements

This work has been part supported by the project "Citizen Observatory WEB" (COBWEB) funded by the European Union under the FP7 ENV.2012.6.5-1 funding scheme, EU Grant Agreement Number: 308513.

References

Alabri A Hunter J (2010). Enhancing the quality and trust of citizen science data. In 2010 IEEE Sixth International Conference on e-Science, 81-88.

Aoki PM Honicky RJ Mainwaring A Myers C Paulos E Subramanian S Woodruff A (2008) Common sense: Mobile environmental sensing platforms to support community action and citizen science. Ubicomp 2008.

Ballatore A Bertolotto M Wilson DC (2013) Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and information systems* (37): 61–81.

Barbier G Zafarani R Gao H Fung G Liu H (2012). Maximizing benefits from crowdsourced data. Computational and Mathematical Organization Theory, 18(3): 257–279.

Brown M Sharples S Harding J Parker CJ Bearman N Maguire M Forrest D Haklay M Jackson M (2013) Usability of Geographic Information: current challenges and future directions. Appled Ergonomics 44: 855–865.

Clow D Makriyannis E (2011). iSpot Analysed: Participatory learning and reputation. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge p 34-43 ACM.

Comber A See L Fritz S Van der Velde M Perger C Foody G (2013) Using control data to determine the reliability of volunteered geographic information about land cover. International Journal of Applied Earth Observation and Geoinformation 23: 37-48.

Díaz P Masó J Sevillano E Ninyerola M Zabala A Serral I Pons X (2012) Analysis of quality metadata in the GEOSS Clearinghouse. International journal of spatial data infrastructures research. 7: 352-377.

Du H Anand S Alechina N Morley J Hart G Leibovici D Jackson MJ Ware M (2012). Geospatial information integration for authoritative and crowd sourced road vector data. Transactions in GIS, 16(4): 455-476.

Fairbairn D Maythm A (2013) Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. International Journal of Geo-Information 2(2): 349–370.

Foody GM, Boyd D (2012) Using volunteered data in land cover map validation: Mapping tropical forests across West Africa. In proceedings of IGARSS, p 2368–2371.

Foody GM See L Fritz S van der Velde M Perger C Schill, C Boyd DS (2013). Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. Transactions in GIS 17(6): 847-860.

Haklay M Weber P (2008). Openstreetmap: User-generated street maps. Pervasive Computing, IEEE, 7(4): 12-18.

Hirth M Hoßfeld T Tran-Gia P (2012). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. Mathematical and Computer Modelling 57: 2918-2932.

Horton JJ Chilton LB (2010). The labor economics of paid crowdsourcing. In Proceedings of the 11th ACM conference on Electronic commerce, p 209-218.

ISO (2002) http://www.iso.org/iso/catalogue_detail.htm?csnumber=26018 ISO

Jackson MJ Rametulla H Morley J (2010) The Synergistic use of authenticated and crowd-sourced data for emergency response. In proceedings of International Workshop on Validation of Geo-Information Products for Crisis Management (VALgEO): 91-99

Kittur A Chi E H Suh B (2008). Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems. P 453-456.

Li D Zhang J Wu H (2012). Spatial data quality and beyond. International Journal of Geographical Information Science 26: 2277–2290.

Martin D Alzua A Lamsfus C (2011). A contextual geofencing mobile tourism service. In proceedings of

Information and communication technologies in tourism: 191-202. Springer Vienna.

Mashhadi A J Capra L (2011). Quality control for real-time ubiquitous crowdsourcing. In Proceedings of the 2nd international workshop on Ubiquitous crowdsourcing, P 5-8. ACM.

Meek S Goulding J Priestnall G (2013) The influence of digital surface model choice on visibility-based mobile geospatial applications. Transactions in GIS. 17(4): 526–543.

Meek S Priestnall G Sharples M Goulding J (2013) Mobile capture of remote points of interest using line of sight. Computers and Geosciences 52: 334-344.

Idris NH Nazri F Said M Iashak M Hashim M Ishmail Z Jackson MJ (2014) Semi-automated metadata detection for assessing the credibility of map mashups through metadata indicators. Forthcoming, 2014 FIG International Congress, Kuala Lumpur, Malaysia,16–21 June 2014

Pawlowicz S Leibovici DG Haines-Young R., Saull, R., Jackson, M., (2011). Dynamic surveying adjustments for crowdsourced data observations. In proceedings Enviroinfo 2011. 5-7 October 2011. Ispra, Italy.

Pourabdollah A Morley J Feldman S Jackson MJ (2013) Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps' Road Network. International Journal of Geo-Information 2(3) 704–728.

Rousell A Jackson M Leibovici DG (2014) Towards automatically identifying the required uncertainty types to describe attributes of geospatial data. Submitted to Journal of Geographical Systems.

See L Steffen F Leeuw J (2013) The Rise of Collaborative Mapping: Trends and Future Directions. International Journal of Geo-Information 2.4: 955–958.

Wang RY Strong DM Guarascio LM (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4): 5-33.

Yang X Blower JD Bastin L Lush V Zabala A Maso J Cornford D Diaz P Lumsden J (2012) An integrated view of data quality in Earth observation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371

Session:
Policy Dimension

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Assessment of the integration of geographic information in e-government policy in Europe

Glenn Vancauwenberghe
KU Leuven - SADL
Celestijnenlaan 200E
Leuven, Belgium
glenn.vancauwenberghe@kuleuven.be

Danny Vandenbroucke
KU Leuven - SADL
Celestijnenlaan 200E
Leuven, Belgium
danny.vandenbroucke@kuleuven.be

Joep Crompvoets
KU Leuven – PGI
Parkstraat 45
Leuven, Belgium
joep.crompvoets@soc.kuleuven.be

Francesco Pignatelli
EC-Joint Research Centre
Via Enrico Fermi 2749
Ispra, Italy
franceso.pignatelli@jrc.ec.europa.eu

Raymond Boguslawski
EC-Joint Research Centre
Via Enrico Fermi 2749
Ispra, Italy
raymond.boguslawski@ext.jrc.ec.europa.eu

Abstract

The integration of geographic information and services in a broader e-government context can be considered as a necessary condition for realising the full potential of Spatial Data Infrastructures (SDIs). In recent years, many European countries have started taking actions and initiatives to integrate geographic information in e-government policy. This paper provides an analysis of these actions and initiatives, focusing on the non-technological aspects, such as the development of strategies, the establishment of coordination structures and the implementation of data policies. The analysis shows that several European countries are aware of the need to integrate geographic information in e-government and are taking different types of actions towards a coordinated and integrated 'information' policy. However, in none of the European countries that were examined is geographic information fully integrated in e-government policy, and in some countries the integration of location information in e-government is even not considered as a priority.
Keywords: e-government, integration, strategies, coordination, data policies

1 Introduction

Over the past ten years significant efforts have been made to improve the access and sharing of geographic information, through the development of Spatial Data Infrastructures (SDIs) at European, national, regional and local level. It is often argued that the benefits of these infrastructures will only be realized once they are in place and are actually being used. A key challenge is to integrate geographic information with other types of information in the different types of processes supporting interactions between public administrations, businesses and citizens. The integration of geographic information and services in a broader e-government context is a necessary condition for realizing the full potential of SDIs. Initiatives to facilitate and promote the use and exchange of geographic information in the public sector will only be successful if they are well connected to e-government [1]. Conversely, initiatives to promote and facilitate the use of geographic information can play an important role in e-government [5]. In that way the relationship between e-government and the use and management of geographic information can be described as symbiotic: while e-government can provide a significant boost to the use of geographic information, the use of this geographic information can be an important enabler for e-government [4].

Despite the clear linkages between geographic information and e-government and the need to integrate both, most policies and initiatives related to the exchange and use of geographic information were originally situated outside the e-government area. While in many European countries the implementation of e-government is managed and coordinated by a separate e-government ministry or agency, the implementation of a coordinated approach on geographic information is often managed by national mapping agencies or Ministries for Environment. This is due to the fact that developments in the geographic information sector were strongly driven by organizations producing data. The involvement of Ministries of Environment on the other hand, is a phenomenon of the past ten years, driven by the emerging INSPIRE initiative aiming to establish an infrastructure for spatial information in Europe [2].

The objective of this paper is to analyze how European countries are taking actions and initiatives to integrate geographic information in e-government policy. The paper focuses on non-technological aspects, such as the development of strategies, the establishment of coordination structures and the implementation of data policies.

2 Methodology

The paper seeks to address the following research question: *What actions are taken at national level in Europe to stimulate and facilitate the integration of geographic information in e-government policy?* In order to answer this question, a survey-based research design was adopted to collect information on the initiatives and actions taken at

country level to support and facilitate the integration of geographic information in e-government.

The survey was targeted at both the public authority officials responsible for e-government (e.g. e-government coordination bodies) and those responsible for geographic information (e.g. INSPIRE National Contact Points) in each EU country. The aim was to collect information on the current status of the use and integration of geographic information in e-government in each Member State from both perspectives: the perspective of the Geographic Information (GI) community and the perspective of the e-government community. This approach also examined awareness levels and involvement of both communities in the use of geographic information in e-government. 23 countries responded to the survey between September and the November 2013. In 12 countries, a representative of the GI/INSPIRE community participated in the survey, in 7 countries the answers were provided by a representative of the e-government community. In 4 countries, both communities completed the questionnaire.

The survey provided information on the actions taken at Member State level to facilitate and coordinate the integration of geographic information in e-government, including development of strategies, establishment of coordination mechanisms, and implementation of data policies.

3 Results

Countries can take actions on several fronts to facilitate the integration of geographic information in e-government. This section analyses the experiences and actions of European countries, focusing on three non-technological aspects: strategies, leadership and coordination, and data policies.

3.1 Strategies

Previous analysis by the European Commission, documented in a series of ePractice e-government factsheets, demonstrated that all European countries have a national e-Government strategy or programme [3]. The degree to which the strategies also focus on geographic information is variable. Of the 23 countries examined in the survey, 4 countries reported that their national e-government strategy does not include any reference to geographic information. In the 19 other countries, the national e-government strategies deal with geographic information in varying degrees. In many cases, the reference to geographic information is relatively limited. For instance, in several strategies attention is only paid to the establishment of the national geo-portal, as a central access point to geographic information. In others, there are only indirect references to geographic information, while none of the objectives or actions in the strategy deal explicitly with geographic information.

In some national e-government strategies a more prominent position has been given to geographic information. Examples of such strategies can be found in Germany, the Netherlands, Switzerland, Finland, Denmark and Sweden. The national e-government strategy of Germany states that available, up-to-date and area-wide reference data are essential for location-based e-government and therefore, spatial data services need to be integrated into e-government applications. The Dutch implementation agenda for e-government services sees geographic information as an important subset of the basic registrations of the country. According to the e-government strategy of Switzerland, geodata should be made available for general use to the authorities of the Confederation, the cantons and communes, the private sector, the public and to academic and scientific institutions in a sustainable, up-to-date, easy-to-use manner, at the required quality and at reasonable cost. In Denmark, shared core data for all authorities, including geographic data, is one of the twelve focus areas of the national e-government strategy, reflecting the strategic objective to integrate geographic information in e-government at all administrative levels. In Finland, the national e-government strategy states that the use of geographic information will improve the quality of services and decision-making and will make public administration more efficient. Therefore, the terms and conditions for governing geographic data should be clear and harmonized and widely used in the public sector. According to the Swedish e-government strategy, the structured management of geographic information is an essential requirement in developing e-services in society. The Swedish public sector must use geographic information that is described in nationally determined references based on international agreements.

In their approach to geographic information, many countries have defined a strategic government framework on geographic information (table 1) Three of the countries examined do not have a strategic document regarding the use of geographic information. In most other countries, a strategy dealing with geographic information in an e-government context is in place. In one of the countries, the strategy only addresses technological issues, in three countries the focus is on organizational issues. The majority of the countries reported that they have a strategy dealing with both organizational and technological issues.

Table 1: Development of a geographic information strategy

Geographic information strategy	Frequency
On organizational and technological issues	14
Only on organizational issues	3
Only on technological issues	1
No	3
No answer/ don't know	2

There are however important differences between these strategies with regard to their content and their focus on the issue of integrating geographic information in e-government. Many national geographic information strategies strongly focus on the development of the national spatial data infrastructure and the implementation of the different components, and pay little attention to the integration and use of geographic information in an e-Government context. Only a few countries have developed a strategy that recognizes the significance of geographic information for realizing the objectives of e-government and defines requirements and actions for raising awareness and extending its use. Interesting examples of strategies dealing with the role of geo-information in e-government can be found in the Netherlands, the United Kingdom, Sweden, Germany and Finland.

One of the key challenges of the geographic-information strategy of the Netherlands was to further develop a geo-information facility in order to give geographic information a prominent place within e-services and e-government. Existing key information facilities, that were created to improve services, enforcement, policy preparation and other processes in government, strongly focused on the creation, management and use of personal data. The same observation is made in the UK Location Strategy, stating that most data in the public sector are related to two aspects: the identification of individuals and companies ('who') and the location of communities, assets, events or environmental conditions ('where'). While the importance of information about citizens and businesses is widely recognized, geographic information is often overlooked. As many areas of policy and service delivery require information on both issues, the UK Location Strategy wants to "complement the focus already being given to 'who' by introducing a parallel focus on 'where'". According to the Swedish SDI strategy, the national spatial data infrastructure should support the development of Swedish e-governance, the Swedish business community and international competitiveness. Improved access to geodata is considered as a precondition for expanded e-governance, and should result in a more efficient administration and a range of new e-services to citizens and businesses.

In Germany, the integration between the national e-government strategy and the geographic information strategy happens at the level of the objectives, as both strategies share the same goals/key objectives: orientation and benefits to citizens, cost-effectiveness and efficiency, transparency, data protection and data security, social participation, future viability and sustainability. Finland is a good example of the shift in focus of the geographic information strategy and activities from data production and availability to the use and integration of geographic information in e-government. According to the Finnish strategy, spatial data services should support people in their everyday activities and during their leisure time, spatial data should be widely used in decision making, should support the participation of citizens, and should be used for managing a large number of functions essential for society.

3.2 Leadership and coordination

Another important dimension in the approaches towards the integration of geographic information in e-government relates to leadership and coordination. Respondents were asked which body or organization was taking leadership in realizing the integration of geographic information in e-government. As shown in table 2, one of the countries indicated a lack of leadership for making geographic information a part of e-government. The other countries have different approaches with regard to the organization(s) responsible for stimulating the integration. Three main groups of countries can be distinguished: countries where the lead is taken by the authority responsible for the Geographic Information policy, countries where the national or regional e-government organization is taking leadership, and countries where leadership is exercised by both the GI and the e-government organization(s). Most of the European countries belong to the first category, and can be considered as 'GI-driven' countries. In almost half of the examined countries, it is the organization or body responsible for GI that takes leadership in the integration of geographic information in e-government. Four countries belong to the second group, as in those countries leadership is provided by the national e-government body or ministry. In six of the countries, leadership in integrating geographic information in e-government is a shared responsibility of the GI and e-government body.

Table 2: Organization leading the integration of geographic information in e-government

Organization taking leadership	Frequency
GI-organization or body	11
E-government organization or body	4
Both organizations	6
Lack of leadership	1
No answer/ don't know	1

Another important organizational dimension of the integration of geographic information in e-government is the establishment of a coordination structure in which members of the e-government community and members of the geographic information community take key decisions. In almost all of the countries, a coordination structure or body involving the e-government community and the geographic information community has been established, and only two countries indicated that they do not have a coordination structure or dedicated body. There are however significant differences in the composition, the role and the tasks of these coordination bodies, which have an impact on their contribution to the integration of geographic information in e-government.

In most countries, consultation and cooperation between representatives of the e-government community and representatives of the GI community takes place in the coordination structure of body that was established to implement the NSDI and/or INSPIRE. In some countries, consultation and cooperation is organized in e-government coordination bodies or groups. In other countries, there is a clear link between the coordination structure for e-Government and the coordination structure for GI/SDI. For instance, in Ireland there is a spatial information subgroup under the Government Offices of the Chief Information Officer. In Sweden, many of the members of the Geodata advisory board are also members of the e-government Delegation.

In Germany, the e-government community and the geographic information community are both represented in the Steering Committee GDI-DE, the coordination and decision-making body for the development of the national SDI. The Steering Committee GDI-DE has been assigned to the IT Planning Council, which constitutes the Central Steering for the Information Technology of the federal and Länder (States) governments. Both bodies, the SC GDI-DE and the IT Planning Council, consist of representatives from federal, provincial and municipal governments. In Switzerland and Germany, joint meetings are regularly organized between representatives of both communities, in addition to consultation and coordination in existing bodies.

3.3 Data policy

A third dimension in which geographic information can be considered as a part of e-government is in the definition and execution of a data policy. The survey focused on two key factors: the presence of an integrated data policy and the presence of a single access point for all data. Table 3 presents the results about the presence of an integrated data policy. In this context, an integrated data policy can be defined as a common data policy that covers all governmental data, i.e. both geographic and non-geographic data of the public sector. It can be concluded that many European countries do not have one common data policy for all their data. From the 23 countries 5 countries indicated that separate policies existed for each dataset or each data provider in their country. In 6 countries, there exists a common policy for multiple datasets, but this policy is limited to only some datasets. While some countries already have an integrated data policy for all geographic data, other countries go further and have an integrated data policy for all their data, both geographic and non-geographic data. In 4 of the countries an open data policy for all data is in place.

Table 3: Presence of integrated data policy at country level

Data policy	Frequency
Open data policy for all data	4
Integrated data policy for all data	4
Integrated data policy for geographic data	4
Common policy for several datasets	6
Each dataset has its own policy	5

Another relevant aspect of the data policy of countries that might stimulate the use and integration of geographic information is the implementation of a single access point for data. Such an access point provides users access to all data sets and services, but also all the relevant information for access and use. Although most countries have at least one access point where several data sets are made accessible, in many countries this access point only provides access to a selection of – geographic - data. In Germany and Poland, all INSPIRE-thematic data are accessible through one single access point, in Switzerland, the Czech Republic and Sweden non-INSPIRE data are also made accessible through this access point. Three countries (Estonia, the United Kingdom and Slovakia) have a single access point for all data, geographic as well as non-geographic data. In the Netherlands, a single access point is under development.

3.4 Discussion

A general conclusion of the analysis is that in none of the European countries surveyed is geographic information fully integrated in e-government policy, in the sense that integration is achieved at the strategic level, at the organizational level and at the level of the data policy. It should however be noticed that several countries (such as Denmark, the Netherlands, Germany and the United Kingdom) are successful in integrating geographic information at several of these levels. Most of these countries already have a well-developed SDI in place, and the challenge for them is to integrate the data and services provided by this SDI in different e-government processes. Many others countries are still in the process of setting up their national SDI. For them, the focus now is on the development and implementation of typical SDI components, and the integration of location information is not considered as a priority

4 Conclusions

The study presented in this paper was designed to explore how European countries are taking actions at national level to stimulate and facilitate the integration of location information in e-government. The focus of this study was on different non-technological measures to align the activities of the GI-community and the e-government community. The evidence from this study suggests that several European countries are aware of the need to bring both communities together and are taking different types of actions towards a coordinated and integrated 'information' policy, considering location information as one of the many types of government information. However, in none of the European countries such a 'fully integrated' information policy already seems to be in place. In many countries the integration of location information in e-government is even not seen as a priority.

A number of important limitations of this study needs to be considered. To begin, the focus of the study was on the non-technological side of integrating location information, although there are also many important technological aspects that should not be neglected. In addition, the study had a strongly explorative character, combining both quantitative and qualitative methods of data collection and analysis, in order to get a first general overview on the state of play in Europe. Additional research is needed in order to gain insight in two crucial areas. First, further investigation is needed to identify the determinants of a certain approach for integrating location information in e-government. Second, and probably most important, further research is needed to better understand which models for integrating location information in e-government are most successful, and lead to an optimal use and integration of location information in e-government services, and better service delivery to citizens and businesses.

At this stage, there doesn't seem to be a "right answer" to how things are organized and different approaches have produced successful results. Often existing organizational responsibilities have played a key role in shaping the way these opportunities are addressed. Nevertheless, there is an increasing trend towards convergence, spurred on by recognition of the contribution that this will make to wider policy objectives of efficiency, growth and better services.

References

[1] M. Craglia, & A. Johnston, A. *Assessing the impacts of spatial data infrastructures: Methods and gaps*. 7th AGILE conference on Geographic Information Science, Heraklion, 2004.

[2] European Commission. (2007). *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).*

[3] ePractice.eu. *eGovernment factsheets*. March 2014. Retrieved from http://www.epractice.eu/en/factsheets/

[4] P. Turner, P. & G. Higgs. The use and management of geographic information in local e-government in the UK. *Information Polity*, *8*: 151 – 165, 2003.

[5] M. Warnest. *A collaboration model for national spatial data infrastructure in federated countries*. Melbourne: University of Melbourne, Department of Geomatics, 2005.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Publishing metadata of geospatial indicators as Linked Open Data: a policy-oriented approach

Diederik Tirry
KU Leuven/SADL
Celestijnenlaan 200E
Leuven, Belgium
diederik.tirry@sadl.kuleuven.be

Ann Crabbé
KU Leuven/SADL
Celestijnenlaan 200E
Leuven, Belgium
ann.crabbe@sadl.kuleuven.be

Thérèse Steenberghen
KU Leuven/SADL
Celestijnenlaan 200E
Leuven, Belgium
therese.steenberghen@sadl.kuleuven.be

Abstract

Geospatial indicators are becoming increasingly important for governments in monitoring and underpinning policy planning and political decision making. Currently, the discovery, viewing and sharing of these indicators is often made possible through geoportals that are developed according the concepts of Spatial Data Infrastructures (SDIs). However, this type of 'business information' exceeds the scope of traditional SDIs that solely focus on the common spatial aspects constituting a generic location context. The concept of an 'augmented' SDI adopting Linked Data principles reveals meanwhile much potential in integrating disparate reference and non-spatial business data but requires a formal revision of underlying standards. In this study we propose an alternative and policy-oriented viewpoint for publishing geospatial indicators as Linked Open Data. Focussing on metadata, we have elaborated a profile of the Data Catalog Vocabulary (DCAT) for describing geospatial indicators, including additional information on the related policy assessments, spatial characteristics, the provenance, and the measurement variables and dimensions of indicators. By implementing the vocabulary in an existing monitoring system it allows us to discuss the benefits and drawbacks of this approach.
Keywords: Linked Open Data, Spatial ontologies, data catalog vocabulary, policy support

1 Introduction

Efficient and effective governance requires reliable knowledge about the current situation, the underlying driving forces, and the consequences and effects of strategic policy plans. For policy makers, the development of integrative monitoring systems is vital in order to process the multitude of information and measure the execution and outcomes of a policy program across time [3]. It is also generally recognized that the use of geospatial indicators in particular can lead to important insights in support of policy and decision making [19].

The 'Spatial Monitor Flanders' and 'Traffic Safety Monitor Flanders' are two examples of monitoring systems that facilitate a multi-level, integrative framework for collecting, publishing and maintaining the most relevant spatial indicators in these policy domains [4,17]. For both monitoring systems the concept of an SDI [8,12] was introduced earlier to connect the scattered and isolated geospatial indicators and create interoperable web services for the discovery, viewing and exchange of relevant information.

Whilst an SDI is intended to enable the access, retrieval and dissemination of geospatial information, the scope of an SDI encompasses solely common spatial aspects constituting a generic location context and therefore does not target specific applications, such as the publication of domain-specific spatial indicators via custom monitoring platforms [18]. When deploying both monitoring systems conform the SDI principles and components, a discrepancy arose between the supply of geospatial indicators and the expectations of policy makers, often less technical in nature. Therefore, the limited scope of SDIs was gradually considered as a major barrier to unlock the full value of geospatial indicators within the policy cycle.

The aim of this research is to bridge the gap between the geospatial community and policy makers by exploring how Linked Open Data (LOD) can be applied in the context of exchanging geospatial and policy-relevant indicators. In this paper we focus in particular on the metadata of geospatial indicators and present a policy-oriented approach for publishing them in the semantic web. The approach relies on the development of a new profile of the W3C Data Catalog Vocabulary (DCAT) to integrate additional metadata elements that are specific and adequate to geospatial and policy-relevant indicators.

The remainder of this paper is structured as follows: first we briefly introduce Linked Data principles and provide an overview of related research. A methodology for developing and applying a vocabulary suitable for describing geospatial indicators is presented in section 3. In section 4 we clarify the benefits and drawbacks of our approach. Conclusions and future research will be discussed in the last section of this paper.

2 Linked Open Data and SDI

The term Linked Data refers to a set of good practices for publishing and connecting structured data in the semantic web, also called the 'web of data' [2]. The notion of Linked Data is underpinned by four core principles introduced by Tim Berners-Lee in his Web architecture note on Linked Data [1]: 1) use Uniform Resource Identifiers (URIs) as reference points, 2) use dereferenceable URIs so that people can look them up, 3) encode the data in the machine-readable Resource

Description Framework (RDF) so they can be queried with the RDF query language SPARQL, 4) include links to other data sources enabling the discovery of related items. As both public and private sector have started to embrace open access and open data policies, the label Linked 'Open' Data (LOD) is now increasingly used referring explicitly to the publication of Linked Data under an open license [7].

LOD provides a new opportunity to study the use and exchange of geospatial data and information in a distributed environment, as well as to re-examine the role of SDIs implementing a service oriented architecture. In addition, the underlying semantic web technologies of LOD offer several benefits to organize the data itself on the Web and thereby using the Web as a global information space.

The use of semantics was first introduced into GIS to enable integration of disparate sources in a seamless and flexible way based on their semantic value and regardless of their representation. The generation and use of ontologies was considered as a method to provide the users with explicit information about the embedded knowledge of the information system thereby enhancing the classification process of various sources of data [6].

Triggered by the success of the LOD community, research recently shifted towards exploring the use of LOD in SDIs. Schade and Cox applied the Linked Data approach to classical SDIs and concluded that SDI concepts and Linked Data principles do no exclude but rather complement each other [15]. Different solutions were proposed to augment SDIs with LOD and improve remaining issues related to cross-community communication and cooperation [16]. At the metadata level Lopez-Pellicer et al. proposed a Linked Data frontend for CSW as a solution for publishing metadata repositories on the Web [10]. Also Reid et al. explored alternative options to publish geospatial metadata as RDF, from 'crosswalking' through well-known vocabularies such as Dublin Core, to RDF generation direct from a relational database [13]. Within the GLUES SDI project, LOD principles and technologies were applied to existing web feature services (WFS) and sensor observation services (SOS) in order to produce RDF representations of service metadata and of respectively features and observations [14]. While the abovementioned studies target individual components, Janowicz et al. presented a shared and integrative Semantic Enablement Layer that comprises a Web Ontology Service for managing ontologies and a Web Reasoning Service for integrating reasoning functionality within SDIs [9].

The concept of augmenting SDIs still faces many challenges, especially towards further elaboration and implementation. First, with regard to geospatial metadata, many of the abovementioned approaches propose well-known vocabularies such as Dublin Core terms. However, these approaches will be partially or fully overtaken if the Open Geospatial Consortium (OGC) and ISO/TC211 committee define themselves a set of Linked Data Vocabularies, hereby following the recommendations of the Delft Report on Linked Data [11]. Secondly, the software infrastructure required to produce and process geospatial Linked Open Data within an augmented SDI is currently limited to stand-alone initiatives and has not reached yet full maturity. Last but not least, most approaches in augmenting SDIs are focussed on leveraging the existing infrastructure in terms of integrating semantics for reference data, unfortunately ignoring the opportunity to establish a common ground for geospatial data and derived products such as monitoring (geospatial indicators) and reporting information.

In summary, the concept of augmented SDIs reveals a lot of potential in connecting the SDI community and the semantic web. However, current implementations are limited to pilots and sharing best practices, waiting on a formal revision of current SDI standards and transformation of existing models to RDF. Consequently, keeping an SDI-based architecture for indicator-based monitoring would impede the publication of geospatial indicators as LOD in the semantic web.

The aim of this research is to explore a new approach for publishing geospatial indicators as LOD, enabling the integration with non-spatial linked data. In the next section we propose a new pragmatic solution to publish geospatial indicators in the semantic web.

3 Methods

For publishing metadata of geospatial indicators, following patterns would be considered according the augmented SDI approach. First, existing metadata can be converted to RDF using an RDF-izer and stored in an RDF triplestore or as static RDF files. Next, the metadata is published on the web using a web server or via a Linked Data interface. Another option is to apply a Linked Data wrapper to access a catalog web service (CSW) and expose a metadata catalogue as Linked Data. Though, both patterns require that all SDI standards fully adopt the Linked Data principles.

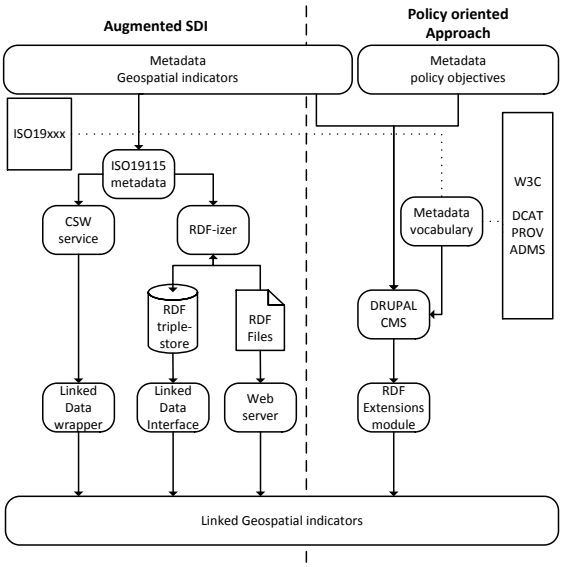
The pattern we propose is inspired by an opposite perspective on integrating metadata of geospatial indicators and Linked Data. Instead of augmenting standards from the SDI we directly select and re-use existing vocabularies that are already well-known and frequently used for describing catalogs within the Linked Data community. By extending these vocabularies with additional metadata elements, we can include information about the spatial characteristics, the policy objectives that are monitored, and the specific measures and dimensions of the geospatial indicator.

The reasoning behind is that geospatial indicators should not necessarily be described applying the ISO19115 standard because derived thematic data are considered out of scope for SDIs. Hence, we could immediately model the metadata starting from existing Linked Data specifications and seamlessly integrate our catalog of geospatial indicators with other data catalogs that are published as Linked Data. Figure 1 presents both patterns.

For the development of a vocabulary we combined a bottom-up approach, based on a use case derived from the Spatial Planning policy in Flanders, with a top-down one, analyzing the relevant semantic vocabularies. The use case helped in identifying the requirements for describing a policy-relevant geospatial indicator, whereas the review of Linked Data vocabularies provided insights into the potential eligibility of existing vocabularies.

Once the vocabulary was elaborated, it was implemented in a geospatial content management system (CMS) in order to have our target audience (i.e. policy makers) use it.

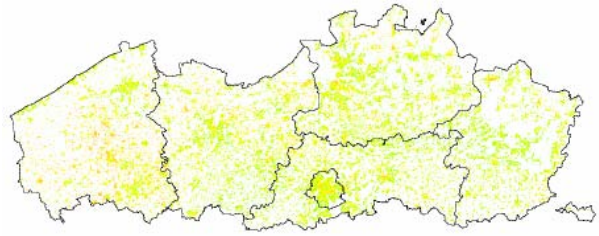
Figure 1: Augmented SDI Linked data publishing pattern (left) compared to policy-oriented approach (right)



3.1 Case study: Multi-level monitoring

Our use case involves the monitoring of the 'Green Infrastructure' for recreational purposes in Flanders. 'Green Infrastructure' is a strategically planned network of natural and semi-natural areas with other environmental features designed and managed to deliver a wide range of ecosystem services[1]. The concept is increasingly recognized by spatial planning authorities as a valuable approach for solving urban and climatic challenges. As the benefits and functions of Green Infrastructure are numerous, we focused on one single application only i.e. the role of Green Infrastructure for recreational purposes. Typical indicators used in a monitoring context are: the general provision of green space, the proximity of green space, the available green space for recreational purposes per person, demand for green space etc... Figure 2 shows an example of a typical geospatial indicator that is monitored at the regional level.

Figure 2: Proximity of green space to place of residence



Source: Natuurrapport Vlaanderen, NARA 2009 [5]

A comparison, however, between a regional and a local monitoring system revealed many semantic differences between the published indicators. We briefly describe the most important types of semantic heterogeneity among the published metadata:

No uniform metadata scheme: Each monitoring system implemented its own metadata schema to describe indicator properties, policy objectives and policy assessments. We determined significant differences in terminology and granularity of meaning.

Use of free-text fields: The ability to provide unstructured information via free-text fields for properties such as provenance, quality and relevance leads to fine-grained knowledge. However, these type of fields are prone to inaccurate information and content mismatch, because it highly depends on the author's competences and willingness to describe these properties in a correct way.

Heterogeneous classifications: Each monitoring system is using its own classification schema to categorize indicators. Whereas the regional monitoring system orders indicators according the concept of ecosystem services, the local monitoring platforms applies their own custom classification schemas. Therefore it is impossible to make a seamless integration between both platforms.

To resolve semantic heterogeneity between the two monitoring platforms we propose the introduction of three semantic components: the definition of an ontology, the adoption of controlled vocabularies and the use of taxonomies.

An ontology allows us to represent the concept of a geospatial indicator in terms of classes and properties that are applied in policy monitoring. The definition of controlled vocabularies enhance the semantic interoperability as the use of free-text is largely reduced to passively recognize a (hierarchical) list of terms as a shared context. Finally, the use of domain-specific taxonomies enables the integration of different types of indicators about the same subject e.g. Green Infrastructure.

3.2 Vocabularies

For the selection of semantic vocabularies we considered the following criteria: 1) a strong user community, 2) stable and open, 3) available in RDF, 4) adequate for our case, 5) unambiguously documented and 6) specific enough to describe indicators in sufficient detail. After a screening of existing Linked Data vocabularies, we concluded that the W3C DCAT[2] vocabulary partially suits our needs. DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. Hence, it supports the monitoring of indicators in different catalogs and by different government bodies. DCAT makes extensive use of terms from the Dublin Core vocabulary, which is well-known and supported by a broad community. Furthermore, it integrates the SKOS[3] vocabulary, enabling the creation of concept schemes for representing policies, structuring

[1] Green Infrastructure (GI) COM/2013/0249 final

[2] http://www.w3.org/TR/vocab-dcat/

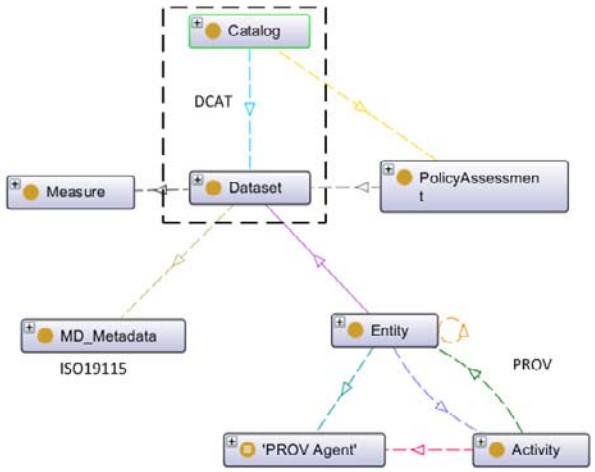
[3] http://www.w3.org/2004/02/skos/

frameworks such as the ecosystem services typology and representing an indicator typology. In the next section we discuss how DCAT can be extended to meet the remaining requirements.

3.3 DCAT-SM vocabulary

In order to meet the remaining requirements of an indicator-based monitoring system, we propose to extend the DCAT ontology and add capabilities to describe policy assessments, spatial characteristics, provenance information, and measurement information as depicted in Figure 3.

Figure 3: Extension of the DCAT vocabulary (extracted from the Protégé Ontology editor)



The result is called DCAT-SM (Data Catalog Vocabulary for Spatial Monitoring). It is developed as a profile of DCAT that includes additional information on:

- Policies: Policy assessments can be described and linked with one or more geospatial indicators. Assessments can be structured according a user-defined taxonomy (e.g. policy objectives) and linked to references and web pages that provide additional details.
- Spatial characteristics: Metadata elements describing the reference system, the resolution and the spatial representation type were extracted from the ISO19115 standard and modelled as RDF classes and properties.
- Provenance: In the context of a monitoring system it is key to understand how geospatial indicators have been calculated. The PROV ontology allows for describing provenance using structured text and/or a graphical representation of the calculation process. Via the entity class of PROV a link can be established to the core reference dataset where the indicator is derived from, avoiding the duplication of metadata elements.
- Measurements: An additional class allows to precisely describe the spatio-temporal dimensions, the thematic dimensions, the measure variables and the units of measurement. This class is indispensable for managing time series and different spatial representations of geospatial indicators. For example, the proximity of green space can be processed and represented using different reference units such as administrative regions or 1km grids.

Besides the definition of classes and properties, the DCAT-SM ontology also prescribes a series of additional classification schemes to better accommodate the Spatial Planning and Road Safety policy context, to adopt the Flemish 'Open Data' licensing framework and to include an indicator typology enabling the distinction between input, output, outcome and impact indicators.

3.4 Vocabulary implementation

The Spatial Monitor Flanders and Traffic Safety Monitor Flanders have been deployed earlier as a geospatial Content Management System (CMS) based on Drupal and integrated with Openlayers, Geoserver and PostGIS to enable geospatial capabilities such as viewing and downloading geospatial indicators.
The DCAT-SM vocabulary has been implemented by transposing each class to a Drupal content type (i.e. pre-defined collection of data types) and each property to a corresponding field type in Drupal. The content type interface allows the users to easily create and edit metadata records of indicators conform the proposed specification.
In addition Drupal has been extended with two existing Drupal modules i.e. 'RDF Extensions' and 'Restful Web Services', hereby providing extra APIs to create RDF representations of metadata records in various serialization formats such as RDF/XML, N-Triples and Turtle.

4 Discussion

Despite the potential of augmented SDIs, the SDI community is struggling with the realization of a common agreed approach for integrating SDIs and Linked Data. A significant issue is the identification of core vocabularies and a methodology how to construct mappings and transform existing metadata (and data) to RDF.
With this study we propose a different approach on the issue of sharing geospatial metadata and purposefully adopted an opposite perspective i.e. integrating Linked Data and SDI by extending Linked Data vocabularies. We try to sum up the most important benefits and drawbacks of this approach. Our approach offers the following advantages :

1. Seamless integration with 'Open Data' Catalogs: Due to the common DCAT vocabulary, catalogs listing geospatial indicators can easily be integrated in the network of emerging 'Open Data' portals.
2. Policy-oriented: The proposed DCAT-SM profile is intended for policy-makers and allows for making indicator-based assessments for any policy domain.
3. Usability: Implementing the vocabulary in an operational CMS exerts two beneficial effects on usability. First, the use of forms allows non-technical users to effortless create metadata records based on the underlying vocabulary. Secondly, the CMS offers high flexibility in

the appearance of policy assessments and geospatial indicators.

4. Accessibility: Additional APIs enable multiple representations (HTML and RDF serializations) and ensure that the content is accessible to different types of users.

Potential drawbacks of our approach are:

1. Isolation from SDIs: the suggested approach is based on the use of the DCAT vocabulary and therefore only partly relies on ontologies derived from ISO19115, disregarding most of the comprehensive schema for describing geographic data. It entails a shift away from SDIs towards the 'open data' community.
2. Narrow scope: In this study an empirical approach to publish metadata as Linked Data has been elaborated, i.e. supporting policy makers with a catalog that structures policy assessments and geospatial indicators. However, a more generic framework including formal extension patterns is indispensable to align and maintain interoperability with current Open Data portals. Ultimately, we consider such a framework as complementary to existing initiatives such as CKAN[4] in order to create catalogs that are fit-for-purpose (e.g. supporting spatial planning policy) and that are embedded in a contentful environment.

5 Conclusions and outlook

The concept of augmented SDIs reveals a lot of potential in connecting the SDI community and the semantic web but requires a formal revision of underlying standards and a transformation of existing models to RDF. In this study we propose an alternative and policy-oriented viewpoint for publishing metadata of geospatial indicators as Linked Open Data. We have established the DCAT-SM vocabulary for describing disparate geospatial indicators, including additional information on the related policy assessments, spatial characteristics, the provenance, and the measurement variables and dimensions. The specification is conceived as a profile of the DCAT vocabulary and is therefore compatible with other catalogs that have applied this RDF vocabulary.

This approach should be considered as a pragmatic and lightweight solution to bridge and integrate spatial thematic data with non-spatial Open Data repositories. With this alternative viewpoint, we also intend to contribute to the challenges on the adoption of Linked Data for geographic information.

Future research will focus on publishing the data itself as Linked Open Data, by exploring the suitability of GeoSPARQL and the RDF Data Cube vocabulary for this specific type of data i.e. geospatial indicators. Simultaneously, we also intend to widen the scope of the current approach in order to establish a more generic and formal framework for describing and distributing geospatial thematic data as Linked Open Data.

[4] http://ckan.org/

References

[1] T. Berners-Lee. Linked Data – Design issues, 2006. Available at http://www.w3.org/DesignIssues/LinkedData.html.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, no. 3, pp. 1–22, 2009.

[3] H. T. Chen. *Practical Program Evaluation: Assess and Improve Program Planning, Implementation, and Effectiveness*. Thousand Oaks, CA: Sage, 2005.

[4] B. Debecker, T. Steenberghen and P. Jacxsens. Spatial Monitor Flanders: Managing spatial data in support of policy making. In *Innovations in Sharing Environmental Observation and Information, Proceedings of the 25th EnviroInfo Conference*, pages 955–966. Shaker Verlag, Marburg, 2011.

[5] M. Dumortier, L. De Bruyn, M. Hens, J. Peymen, A. Schneiders, T. Van Daele & W. Van Reeth (red.). Natuurverkenning 2030. Natuurrapport Vlaanderen, NARA 2009. Mededeling van het Instituut voor Natuur- en Bosonderzoek, INBO.M.2009.7, Brussel.

[6] F. Fonseca, M. Egenhofer, P. Agouris and C. Câmara. Using Ontologies for Integrated Geographic Information Systems. Transactions in GIS 6(3), pp. 231-257, 2002.

[7] C.P. Geiger and J. von Lucke. Open Government and (Linked) (Open) (Government) (Data), *Journal of e-Democracy and Open Government*, vol. 4(2), pp. 265–278, 2012

[8] R. Groot and J. McLaughlin. *Geospatial Data Infrastructures.* Oxford, Oxford University Press, 2000.

[9] K. Janowicz, S. Schade, A. Bröring, C. Keßler, P. Maue, and C. Stasch. Semantic Enablement for Spatial Data Infrastructures. *Transactions in GIS* 14(2), Blackwell Publishing, pp. 111-129, 2010.

[10] F. J. Lopez-Pellicer, A.J. Florczyk, W. Rentería-Aguaviva, J. Nogueras-Iso and P.R. Muro-Medrano. CSW2LD: a Linked Data frontend for CSW. In *II Iberian Conference on Spatial Data Infrastructures*, Institut Cartogràfic de Catalunya, 2011

[11] F. J. López-Pellicer, L.M. Vilches-Blázquez, F.J. Zarazaga-Soria, P.R.. Muro-Medrano, and O. Corcho. The Delft Report: Linked Data and the challenges for geographic information standardization. *Jornadas Ibéricas de Infraestructuras de Datos Espaciales* (JIIDE 2011), Barcelona, 2011.

[12] D. Nebert. *Developing Spatial Data Infrastructures: The SDI Cookbook.* Version 2.0, Global Spatial Data

Infrastructure Association, Technical Working Group Report, 2004.

[13] J. Reid, W. Waites and B. Butchart. An Infrastructure for Publishing Geospatial Metadata as Open Linked Metadata. In Proceedings of AGILE 2012 International Conference on Geographic Information science, Avignon, 2012

[14] M. Roth and A. Bröring, editors. *Linked Open Data in Spatial Data Infrastructures*. Available at https://wiki.52north.org/pub/Projects/GLUES/2012-09-10_LoD_SDI_White_Paper_MR_AB.pdf

[15] S. Schade and S. Cox. Linked Data in SDI or How GML is not about Trees. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science - Geospatial Thinking*, Guimarães, 2010.

[16] S. Schade, C. Granell, L. Díaz. Augmenting SDI with Linked Data. In *Proceedings of the Workshop on Linked Spatiotemporal Data*, GIScience, 2010.

[17] D. Tirry and T. Steenberghen. Towards a semantic-driven spatial monitoring framework for Road Safety. *25th ICTCT workshop*, Hasselt, 2012.

[18] K. Tóth, C. Portele, A. Illert, M. Lutz and M. N. De Lima. *A conceptual model for developing interoperability specifications in Spatial Data Infrastructures.* JRC Reference reports, Ispra, 2012.

[19] I. Williamson, A. Rajabifard and M.A.F. Feeney, editors. *Developing Spatial Data Infrastructures: From Concept to Reality.* Taylor and Francis, London, 2003.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Exploring the market potential for geo-ICT companies in relation to INSPIRE

Glenn Vancauwenberghe
KU Leuven - SADL
Celestijnenlaan 200E
Leuven, Belgium
glenn.vancauwenberghe@kuleuven.be

Piergiorgio Cipriano
Sinergis
Via del Lavoro 71
Casalecchio di Reno, Italy
piergiorgio.cipriano@sinergis.it

Max Craglia
EC-Joint Research Centre
Via Enrico Fermi 2749
Ispra, Italy
massimo.craglia@jrc.ec.europa.eu

Cameron Easton
Gistandards
Swift Bank 1
Hamilton, United Kingdom
cameron@gistandards.eu

Giacomo Martirano
Epsilon Italia
Via Pasquali 79
Mendicino, Italy
g.martirano@epsilon-italia.it

Danny Vandenbroucke
KU Leuven - SADL
Celestijnenlaan 200E
Leuven, Belgium
danny.vandenbroucke@kuleuven.be

Abstract

The implementation of INSPIRE can bring new and interesting business opportunities to European geo-ICT companies. Until now, little information has been available on the participation of geo-ICT companies in the implementation of INSPIRE. This paper seeks to explore the market potential for geo-ICT companies in relation to INSPIRE, presenting the results of a large-scale survey among geo-ICT companies in Europe. The paper shows that the majority of geo-ICT companies in Europe is not actively involved in the implementation of INSPIRE. Having knowledge and understanding of the technical details of INSPIRE seems to be a key requirement for companies to get involved in INSPIRE. Companies that fulfil this requirement and have supported public authorities in implementing INSPIRE, have experienced an impact of INSPIRE on their innovative performance.
Keywords: INSPIRE, geo-ICT sector, impact of INSPIRE, innovation

1 Introduction

The INSPIRE Directive establishes an Infrastructure for Spatial Information in Europe to support Community environmental policies, and policies or activities which may have an impact on the environment [4]. INSPIRE is based on the creation, operation and maintenance of infrastructures for spatial information established and operated by the 28 Member States of the European Union plus Switzerland, Norway and Iceland, addressing 34 spatial data themes related to environmental applications.

Making data available, according to INSPIRE standards, requires specific skill sets that sometimes are difficult to find in public authorities. The management of this content represents an opportunity for enterprises active in this sector, and small and medium-sized enterprises (SMEs) in particular. Offering their services and products, geo-ICT companies and SMEs can help governments in the implementation of the requirements imposed by INSPIRE. It is expected that the technical skills and organizational flexibility of SMEs can effectively support the various institutions and stakeholders directly involved in the various commitments related to the implementation of INSPIRE. Due to legal requirements, the INSPIRE implementation can become the entry-point for crucial business opportunities, opening new or reinforcing existing perspectives.

In order to explore the market potential for geo-ICT companies in relation to INSPIRE and to define the obstacles for geo-ICT companies to enter this market, insight is needed in the characteristics, knowledge and activities of geo-ICT companies in Europe, especially related to the implementation of INSPIRE. In the context of the European smeSpire project a large-scale survey was organised among geo-ICT companies in Europe. Making use of the results of this survey, this paperthis paper analyses in detail the involvement of the European geo-ICT sector in INSPIRE in order to provide some valuable recommendations on how the participation of the geo-ICT sector in Europe to INSPIRE can be promoted and stimulated.

2 The geo-ICT sector in Europe

Little information and data is available on the overall European geo-ICT sector. However, some studies focus on the geo-ICT sector in one single Member State. Castelein W.T. et al made an analysis of the Dutch Geo-ICT sector in 2008. Their analysis showed that in that year, the Dutch private Geo-ICT sector had a turnover of € 900 million from geo-information products and services to which 9977 employees contributed [2]. The private sector was responsible for 66% of the total "geo" workforce and 64% of the overall geo-information economic value. Geo-ICT accounted for 3.64% of the total number of ICT employees and 1.04% of the overall number of ICT companies. The most important domain of the private geo-ICT sector, with a total turnover of € 297 million, was measuring, collecting and storing geographic data.

For several years, AGORIA, the Belgian federation for the technology industry, has assessed the Geo-ICT sector in Belgium [1]. Most recent figures show the Belgian geo-ICT sector comprises approximately 60 companies, generating a total annual turnover of more than € 335 million, and offering jobs to an estimated 1850 employees. The UK Location Market Survey 2012 provides an assessment of both the current size and future directions of the UK Market for Location Information Products and Services [3, 5]. The estimate for location related software, professional services, data and hardware in 2012 is €1.49 billion. The authors also predict continued growth at a modest 1 to 2% in real terms over the following 3 years.

Information on the size of the German geo-ICT sector has been provided by MICUS [6]. In the year 2000 the market volume amounted to €1 billion, and by 2007 this had increased by 51% to just over €1.5 billion. According to the report, the geo-business market can be classified into three main sectors: navigation and mobile services, planning and documentation systems and geo-marketing. Notably, in the navigation sector the volume of sales more than doubled between 2000 and 2007, from €350 million to €728 million.

When comparing and generalizing the results of these studies, it should be noticed that there is no generally agreed definition of the term 'geo-ICT sector' and most existing studies and policy documents use their own definition. Castelein et al state that "the geo-information sector works with location specific (x,y,z) information or services". Within the, four areas of activity can be identified: 1) measuring, collecting and storing of data about geo-objects; 2) processing, editing, modelling, analyzing and managing that data; 3) presenting, producing and distributing the data; and 4) advising, educating, researching and communicating about processes and use of geo-information products and services [2]. According to AGORIA the geo-ICT sector deals with information related to geographical location by providing solutions in the area of the Geographical Information Systems (GIS) which are designed to gather, store, process, analyze, manage, organize, present and diffuse all types of geographical data [1]. ConsultingWhere refers to "economic activities where geographic information is the main driver of the application, service or system component" [3]. Some of the existing definitions include a clear reference to the ICT sector, highlighting that the ICT sector is frequently considered a main reference sector for private companies dealing with geographic information and geomatics. In the context of this paper, the (private) geo-ICT sector was defined as all companies directly or indirectly involved in the creation and publishing of spatial data and/or in more traditional GIS/geo-location based activities.

3 Methodology

In order to explore the market potential for geo-ICT companies in relation to INSPIRE, this paper addresses the following three research questions: *1) To what extent do European geo-ICT have knowledge and awareness of INSPIRE? 2) How are European geo-ICT companies currently involved in the implementation of INSPIRE? & 3) Does INSPIRE have an impact on the innovative performance of geo-ICT companies in Europe?* In order to better understand the geo-ICT sector in Europe and its involvement in INSPIRE, a large scale survey among geo-ICT companies in Europe was organized between November 2012 and August 2013. The aim of this survey was to gain insight in the characteristics, the activities and the skills and knowledge of geo-ICT companies in Europe. The survey questionnaire consisted of three main parts and included more than 30 questions. The first part of the survey focused on the general characteristics of geo-ICT companies in Europe. In this part, information was collected on the location of the company, the year of foundation, the number of employees, the geospatial activities of the company, etc. The second part of the survey deals with the knowledge and skills of companies related to geo-ICT and INSPIRE. Here, information was collected on the awareness about INSPIRE, the knowledge on INSPIRE and the execution of INSPIRE-related activities in the organization. The third part of the survey was dedicated to the impact of INSPIRE on the organization, with specific attention to the issue of innovation.

All project partners of the smeSpire project were involved in distributing the survey, and inviting geo-ICT companies to participate. In some countries the INSPIRE contact points and the national GI-association supported the distribution of the survey. The survey was completed by 299 geo-ICT companies. In terms of workforce, almost all participating geo-ICT companies fall within the category of 'small and medium-sized enterprises'. 59.4% of the participating companies were even 'micro enterprises', with less than 10 employees. As 31.5% of the companies were 'small' (between 10 and 50 employees), only few medium-sized (between 50 and 250 employees) or large companies were involved in the study. Also in terms of the annual turnover, almost all participating geo-ICT companies were 'micro', having a turnover of less than €1million per annum. 24% of the involved companies had an annual turnover between €1million and €10million, in 3% of the companies the annual turnover was even higher than €10million.

Besides the number of employees and the annual turnover also other background information was collected on the characteristics of the companies. 90% of the companies were created between 1988 and 2008, of which 34% during the 1990s and nearly 12% after 2000. More than 15% of the companies that participated in the survey were part of a larger group. The market level of geo-ICT companies is mainly sub-national, with almost half of the companies surveyed (46%) indicating their primary market is local, and their secondary market (41%) is national. The public sector is the principal customer for European geo-ICT companies representing more than half of the business for 63% of the companies. For the large majority of companies (85%) customers are mainly public authorities within their own country, covering both national and local administrations. 32% of the companies were involved in one or more EU co-funded projects in 2011. More than half of the companies analyzed indicated their core business were geospatial activities, meaning that more than 80% of their annual turnover comes from products or services strictly related to geographic information. 39% of the companies considered themselves primarily as 'users' of spatial data, in 27% of the companies the primary activity was the development of client applications. 20% of the companies

were primarily involved in data modelling and/or the transformation of spatial data.

4 Results

4.1 Awareness and knowledge of INSPIRE

Of the participating companies, 69% indicated to be aware of the INSPIRE Directive. This means that INSPIRE is not known by 31% of the companies. In addition to the general awareness of INSPIRE, also the knowledge about different aspects of INSPIRE was measured. However, only companies that were aware of INSPIRE, were asked to report on their knowledge of different INSPIRE aspects. As can be seen from table 1, the general aspects of the Directive are well known, but companies are less familiar with the more detailed technical aspects.

Almost half of the geo-ICT companies in Europe that are aware of INSPIRE indicated to have high or even very high knowledge of the general objectives (46%) and the main principles of INSPIRE (44%). Knowledge on the conceptual framework, metadata regulation, data and service sharing regulation and interoperability of data and services regulation was relatively lower, and especially the regulations on network services regulation and on the monitoring and reporting obligations were less known.

Table 1 Knowledge on different aspects of INSPIRE

	(Very) low	Medium	(Very) high
Objectives of INSPIRE	36%	18%	46%
Main principles	37%	19%	44%
Conceptual framework	42%	22%	36%
Metadata regulation	44%	20%	36%
Data and Service Sharing regulation	45%	20%	35%
Network services regulations	50%	19%	31%
Interoperability of data and services regulations	48%	17%	35%
Monitoring & reporting obligations regulation	54%	22%	24%

4.2 Involvement in INSPIRE

Only 34% of the participating geo-ICT companies were somehow involved in INSPIRE activities. Most of these companies were involved in INSPIRE working as a contractor for public authorities implementing INSPIRE (20%). Few companies were involved in the development and implementation of INSPIRE as a member of a Spatial Data Interest Community of INSPIRE (10%) or as an expert within one of the Thematic Working Groups (5%).

With regard to the INSPIRE activities that were performed or the INSPIRE components that were developed, the geo-ICT companies were mainly involved in data modelling (26%), the development of view services (26%) and the implementation of metadata catalogues (21%). Activities in which the current involvement of INSPIRE was still low, were setting up test suites (12%), and performing schema transformations (9%).

An explanation of the relatively low involvement of geo-ICT companies in INSPIRE implementation can be found in the fact that many companies consider themselves as 'spatial data users'. This means these companies are not directly involved in the implementation of INSPIRE, but will rather make use of the data and services provided by INSPIRE. The relevance of INSPIRE to many of these companies is demonstrated in the INSPIRE data themes, and the extent to which the activities of geo-ICT companies are related to these themes.

Table 2 shows the main INSPIRE data themes to which the activities of European geo-ICT companies are related. The most interesting data themes for these companies are land use (57%), cadastral parcels (50%), co-ordinate reference systems (50%), land cover (47%) and buildings (46%). Many of the INSPIRE data themes thus are relevant to a relatively large group of geo-ICT companies in Europe.

Table 2 Relevance of INSPIRE data themes

	Companies active in theme
Land use	57%
Cadastral Parcels	50%
Co-ordinate reference systems	50%
Land cover	47%
Buildings	46%
Orthoimagery	45%
Elevation	44%
Transport networks	43%
Addresses	42%
Utilities and government services	42%

4.3 Impact of INSPIRE

The last part of the survey focused on the impact and innovative potential of INSPIRE to geo-ICT companies. Companies were asked to report which changes already occurred in their organization due to the INSPIRE Directive and which changes they expected to occur in the near future. It can be seen from the data in table 3 that INSPIRE already had an impact on many geo-ICT companies, and this impact is expected to increase in the following years. The current impact of INSPIRE is mainly related to the introduction of new or significantly improved products and services (42%) and the introduction of new or improved methods of producing (33%). In the future, the responding companies also expected to see an impact in the emerging of new customer markets (72%), in addition to the introduction of new or improved products and services (74%).

Table 3 Occurred and expected changes due to INSPIRE

	Occurred changes	Expected changes
New or improved products/services	42%	74%
New or improved methods of producing	33%	67%
New customer groups/geographic markets	31%	72%
Product/service delivery in less time or lower cost	28%	66%

4.4 Discussion

The results of the study reported in this paper provide insight in the meaning of INSPIRE for the geo-ICT companies in Europe. One of the most interesting findings of this study was that almost one in three geo-ICT companies in Europe is not aware of INSPIRE. Companies that are aware of INSPIRE have good knowledge of the general objectives and principles of INSPIRE, while the more technical details of INSPIRE are less known. Based on these findings, the geo-ICT companies can be divided into three equal groups: a group of companies that is not aware of INSPIRE, a group of companies that knows the general aspects of INSPIRE and a group of companies that has more advanced knowledge on INSPIRE. It is especially the latter group that is directly involved in the implementation of INSPIRE, mainly working as a contractor for public authorities. The other two groups are currently standing on the side lines while public administrations are implementing INSPIRE.

As one might expect, it is the group of companies that is actively involved in the implementation of INSPIRE that also has experienced the impact of INSPIRE on their innovative performance. The most common change caused by INSPIRE is the delivery of new products and services by companies. However, until now, the majority of European geo-ICT companies did not experience any impact of INSPIRE on their own activities. Looking at it from the positive side, it can be noticed that most of the companies expect to see an impact of INSPIRE in the near future. As many of the European geo-ICT companies primarily are data users, there is a great expectation that INSPIRE will contribute to growth in the future by making data and services available to businesses and allowing them to create added value services.

5 Conclusions

This paper has investigated the involvement of European Geo-ICT companies in the implementation of INSPIRE. The results presented in this paper show that a relatively small group of geo-ICT companies in Europe is actively involved in the implementation of INSPIRE. These companies have more than basic knowledge and competences on different aspects of INSPIRE. Due to their active participation in INSPIRE, they are able to turn the INSPIRE European Directive into business opportunities for their company, leading to growth and innovation. However, many European geo-ICT companies currently are not engaged in the implementation of INSPIRE. Some of these companies are even not aware of INSPIRE.

Changing this situation requires commitment and efforts of both geo-ICT companies and public organizations. Geo-ICT companies need to build up a critical mass on INSPIRE, focused on real needs and requirements of public administrations. Companies should also get more involved in INSPIRE debates, and reflect and communicate about how they can help administration in fulfilling the requirements of INSPIRE. For public administrations, the challenge is not only to take advantage of the knowledge and competences of private companies for the implementation of INSPIRE, but especially to provide companies the opportunity and stimulate them to create add value products and services on INSPIRE data and services.

References

[1] AGORIA. *Analysis of the Belgian GEO-ICT sector*. AGORIA internal report, 2012.

[2] Castelein, W.T., Bregt, A., & Pluijmers, Y. The economic value of the Dutch geo-information sector. *International Journal of Spatial Data Infrastructures Research*, 5:58-76, 2010.

[3] ConsultingWhere. Assessing the UK market for GI. *GISProfessional*, 29:20-21, 2009.

[4] European Commission. (2007). *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).*

[5] Masser, I., & Waters, R. *Estimating the impact of INSPIRE implementation on the UK commercial sector.* INSPIRE Conference, Istanbul, 2012.

[6] MICUS. *European Legislation as a driver for German Geobusiness.* MICUS Management Consulting GmBH. Düsseldorf, 2010.

Session:
Data Mining

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Analysing spatiotemporal patterns of antibiotics prescriptions

Luise Hutka and Lars Bernard
Technische Universität Dresden
Professorship of Geoinformation Systems
Dresden, Germany
luise.hutka@tu-dresden.de,
lars.bernard@tu-dresden.de

Abstract

The emergence of antibiotic resistances due to antibiotic residues in urban sewage systems is becoming an increasingly important issue. This paper presents a model for the spatiotemporal analysis of antibiotic inputs to derive spatiotemporal distribution patterns which are the basis for later predictions of future antibiotic inputs into the sewer system. To identify spatiotemporal distribution patterns of antibiotic prescriptions data statistical and GIS methods like time series and spatial cluster analysis are used. In order to find possible interrelationships the prescription data is combined with other influencing parameters (e.g. cases of respiratory infections) and tested for statistical correlations. Results show a pronounced seasonal course for three antibiotics of the macrolide group which also show high correlations with cases of respiratory infections in the study area. Further, results show that weekly data of respiratory infections by *Google Flu Trends* may be used as predictor variable to derive forecasts of future antibiotic inputs into the sewer system.

Keywords: antibiotic prescriptions, spatiotemporal pattern recognition, drug residues, correlation analysis

1 Introduction

Drug residues in sewage are an important issue in the context of the objectives of wastewater treatment [9, 12, 13]. In general, it is assumed that an increased input of antibiotics into the environment promotes the formation of antibiotic-resistant bacteria. If a pathogen is resistant to a particular antibiotic, taking this antibiotic in case of an infection with the pathogen is ineffective. In the end, the increasing antibiotic resistances together with the decreasing development of new antibiotics in recent years may lead to more and more antibiotics becoming ineffective, with the result that infectious diseases could spread again [19]. Thus, there are numerous studies on the efficiency of various procedures regarding the specific behaviour of antibiotics and resultant antibiotic resistance [1, 16, 17].

The project ANTI-Resist[1] [5] researches the release of antibiotics and potentially related appearance of antibiotic resistances in the urban sewage system of the city of Dresden. As a long term objective the project is meant to support the design of strategies to reduce the formation of antibiotic resistances in urban wastewater. The project focuses on the development of corresponding monitoring and warning systems. Within the project various aspects related to antibiotic fluxes and transports in urban wastewater systems are considered. First, the antibiotic prescriptions of medicines are investigated. Second, models are being developed to describe their release and transport within the sewage system and third, the related emergence of antibiotic resistances within the sewer and the water treatment facilities are studied using different measurement and observation methods.

This paper describes the approach of spatiotemporal analysis and modelling of antibiotic prescriptions using statistical tools and GIS. The paper starts in describing the study design to analyse spatiotemporal patterns of antibiotics input into the sewage system using existing medical prescription data and further geodata. The remainder focuses on discussing the current results and on detecting appropriate input variables – e.g. from crowd-sourced data – for a model to predict antibiotic release into the sewer system.

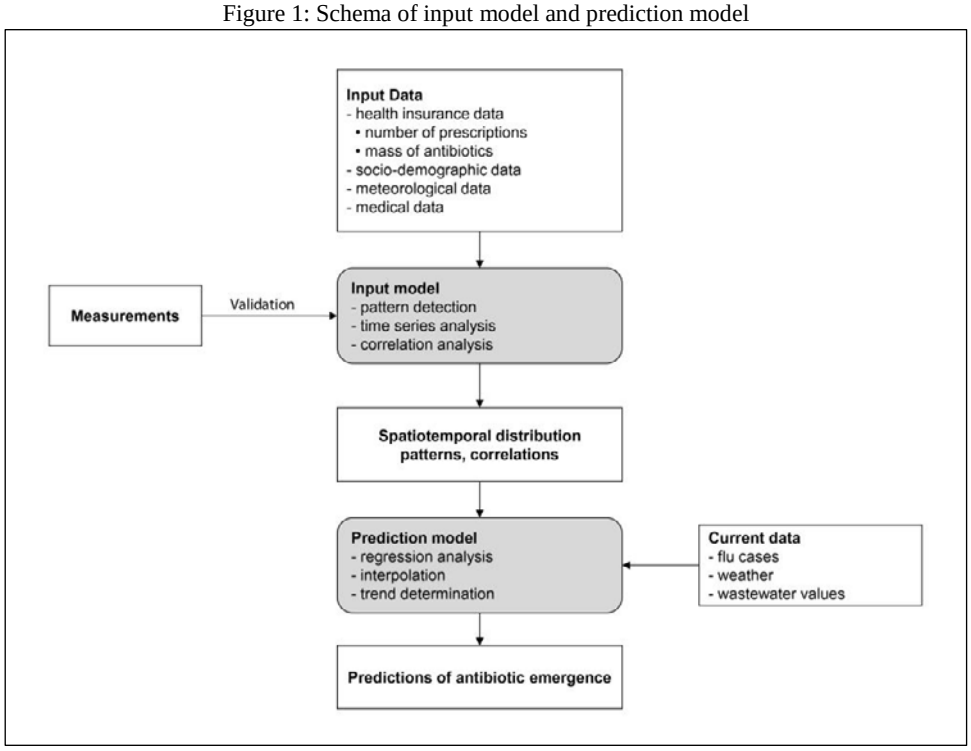
2 Spatiotemporal Analysis of Antibiotic Medications

Various studies demonstrate GIS to be a powerful tool for analysing spatial and temporal distribution patterns of drug-related health data. Cheng et al. [2] used local spatial association statistics to examine the geographic variation of cardiovascular drug-prescribing patterns in Taiwan. Modarai et al. [15] performed a local Moran's I analysis on annual opioid prescription sales aggregated by 3-digit zip codes and correlated this data with official data on opioid overdoses. As in the previous studies, also in the present work local Moran's I analysis is performed on prescription data, but in applying a higher temporal and spatial resolution as the used antibiotic prescription data are available at a weekly/monthly resolution for 8 years and on an urban district level. A related work on antibiotic prescriptions was published by Kern et al. [11] who studied the regional variation in outpatient antibiotic use within Germany. In contrast to the work presented here, Kern et al. analysed prescription data for only one year and on the large-scale federal state level, further they did not apply GIS-based spatial analysis.

The ANTI-Resist project follows a twofold approach in analysing spatiotemporal patterns of antibiotics emergence and related processes in urban wastewater: (1) Sewer measuring campaigns and related laboratory studies are

[1] http://anti-resist.de/

Figure 1: Schema of input model and prediction model



conducted to better understand the actual antibiotic fluxes and the genesis of antibiotic resistance within the sewer system and (2) data driven (statistical) models are getting designed to serve as best estimated guesses on where antibiotics are being released and how they are transported through the sewage system. Figure 1 sketches the conceptual frame for these models and consists of two components:

1. An *input model* has been designed to analyse historical prescription data, socio-economic data and environmental data to identify typical spatiotemporal patterns of the antibiotic medications, to estimate related antibiotic fluxes into to the sewer system and to evaluate these estimates against different measurements within the sewer system.
2. The input model then serves as the basis for a *prediction model* to estimate the future release and fluxes of antibiotics into the sewer system. The *prediction model* shall serve to alert the urban waste water treatment and environmental agencies about the potential occurrence of antibiotic peaks in the sewage and the released waste water. As the treatment of antibiotics is not part of the operational waste water treatment these alerts could also trigger related specific measurements in the sewage plant.

The main objective of the input model is to use spatial statistical tools to investigate the variation of antibiotic inputs in time and within the urban districts of Dresden as well as correlate it with other influencing factors that might have an impact at the variation of the antibiotic prescriptions. This paper focuses the design and results of the *input model* and will discuss some initial design ideas for the prediction model.

2.1 Input Data

Basis for the input model are data about the ambulant antibiotic prescriptions that have been provided by one of the German compulsory health insurances (*AOK PLUS*). These data are available in a weekly and monthly temporal resolution for the period from 2005 to 2012. The data are provided for the 64 urban districts of the city of Dresden and aggregated into three age groups: 0 to 14 years, 15 to 64 years, 65 and more years. The following antibiotic substances of various active ingredient groups are examined: amoxicillin, azithromycin, cefuroxime, ciprofloxacin, clarithromycin, clindamycin, doxycycline, levofloxacin-ofloxacin, phenoxymethyl penicillin, roxithromycin and sulfamethoxazole-trimethoprim.

Several issues arose when analysing the prescription data from the health insurance. First, the supplied data comprise only the patients of this single insurance company, which represent about 41% of the population of Dresden. So the prescription data is extrapolated to the total population of Dresden. Second, due to medical data protection issues prescription amounts between 1 and 3 are anonymised. In the calculations this problem is handled by setting all anonymised values to the minimum amount of 1. Third, the data can only be delivered with a delay of at least one year. Consequently up-to-date official prescription data are not available for the studies. This fact is seriously hampering the prediction of future antibiotics release into the urban sewer system. To overcome the latter problem it was necessary to identify data of other influencing factors that are related to antibiotic prescriptions and would eventually serve as a proxy for a prediction of antibiotic medication. The identified major influence factors are presented in the following.

Several studies have shown that there is a correlation between respiratory infections and antibiotic prescriptions. On the one hand for the majority of treated cases of respiratory diseases antibiotics are prescribed. This is evident especially during the annual cold waves and winter flu season, which regularly are accompanied by a significant increase of antibiotic prescriptions [4, 8, 20]. On the other hand often a secondary bacterial infection follows a flu infection, as the organism is already weakened due to the fight against the viruses. Therefore bacteria can more easily lead to further infections that are then often treated by the use of antibiotics – where partially antibiotic prescriptions are filled prophylactically, without bacterial caused symptoms being already present [10, 18]. For these reasons contemporary available representative data about current cases of respiratory diseases have been considered as potential proxies for the input model. Such data is provided by the flu trends portal from Google[2]. Ginsberg et al. [6] have analysed billions of Google search requests and determined that there is a high correlation between the frequency of particular search terms and the actual number of patients with influenza-like symptoms at a time. For the validation of their results they used historical data of traditional influenza surveillance systems. For Germany *Google Flu Trends* provides the weekly cases of respiratory infections per 100,000 inhabitants at a Federal State level from 2003 up to the current week [7].

Other influencing factors that have been identified and incorporated into the input model are meteorological (temperature, precipitation, etc.) and sociodemographic (population, employment structure) parameters.

For the meteorological parameters the *Regional Climate Information System for Saxony, Saxony-Anhalt and Thuringia* (ReKIS)[3] is used. This database contains – for the city of Dresden – data for 3 climate stations and 3 precipitation stations in a daily resolution for the years 1961 to 2013. At the most comprehensive station about 16 parameters are determined.

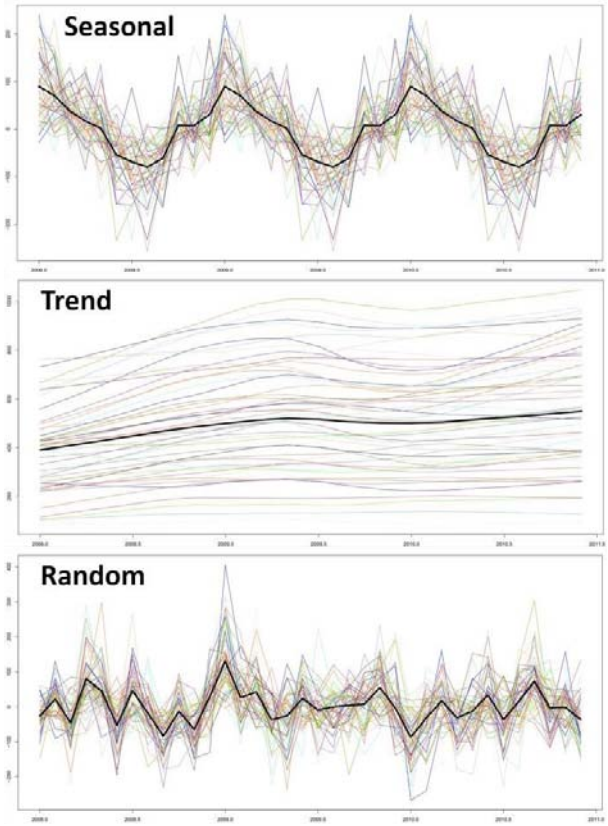
For the sociodemographic parameters data from the Dresden statistics office is used[4]. It provides yearly data at an urban district level about the population, discriminated by sex and nine age groups and data on employees and unemployed persons as relative proportions of the relevant age group. The health insurance company *AOK PLUS* provided data on the age structure of their patients in 5 age groups at the urban district level for the year 2011. Most of the input data described here can be explored via the ANTI-Resist Geoportal[5].

2.2 Methods

In a first step the antibiotic prescriptions have been modelled regarding their temporal and spatial distribution using various statistical and GIS-based methods.

The temporal analysis has been carried out by time series analysis in R using the STL function, which decomposes a time series into three components: seasonal variation, long-term trend and random noise (remainder component) [3]. Thus, the temporal variation of the prescriptions of the individual antibiotic substances in the overall city as well as in the individual urban districts could be examined as to whether there are certain long-term trends and periodic seasonal patterns. An example of the result of such a time series analysis is shown in Figure 2.

Figure 2: An example for the resulting components of a time series analysis in R using the STL function (sum of all antibiotic substances, 2008 – 2010)



ArcGIS (ESRI ArcGIS 10.2) with its geostatistical methods for pattern recognition has been used to analyse the spatial distribution of the antibiotics prescriptions within the Dresden urban districts. ArcGIS offers several functions for the analysis of local spatial patterns: *Hot Spot Analysis (Getis-Ord Gi*)*[6] and *Cluster and Outlier Analysis (Anselin Local Morans I)*[7]. Figure 3 and 4 show examples of the results of the respective cluster method. The Hot Spot Analysis results in statistically significant clusters of similarly high values (Figure 3 - red) or low values (Figure 3 - blue) for different confidence levels. In contrast, the result of the Cluster and Outlier Analysis shows not only statistically significant clusters of high values (Figure 4 - red) or low values

[2] http://google.org/flutrends/

[3] http://www.rekis.org

[4] www.dresden.de/de/02/06/auskunft/medien/atlas.html

[5] http://antiresist.dyndns.org/client/

[6] http://resources.arcgis.com/en/help/main/10.2/#/Hot_Spot_Analysis_Getis_Ord_Gi/005p00000010000000/

[7] http://resources.arcgis.com/en/help/main/10.2/#/Cluster_and_Outlier_Analysis_Anselin_Local_Moran_s_I/005p0000000z000000/

(Figure 4 - blue) but also districts of high values surrounded by low values (Figure 4 - yellow) and vice versa (Figure 4 - white). As the Cluster and Outlier Analysis seemed more meaningful and as it also provides hints to outlier districts, this method has been selected.

In this way, the spatial and temporal prescribing patterns of antibiotics in Dresden and their changes over time have been determined. In a next step possible explanations for the occurrence of specific distribution patterns should be identified. Therefore the antibiotic prescription data are analysed to search for statistical correlations (next section) considering the above mentioned influencing factors. The correlation analyses are mostly performed via simple linear Pearson correlation (using SPSS Statistics 21).

The input model still needs to be validated with actual measurements in the sewer system made within the project. However, a successful validation study strongly depends on the identification of appropriate measuring points that can be defined as being comparable with the prescription data at the urban district level.

Figure 3: Hot Spot Analysis of azithromycin prescriptions in January 2009

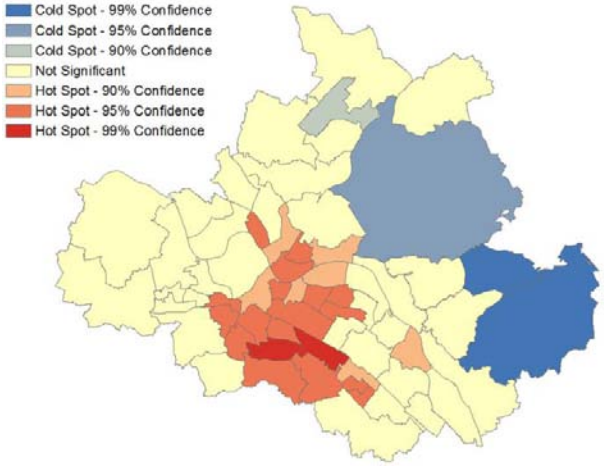
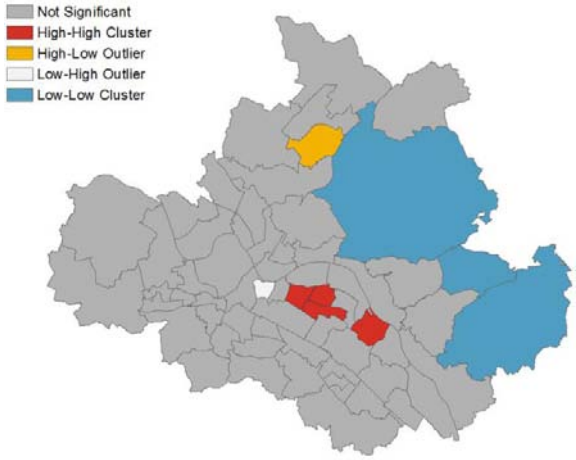


Figure 4: Cluster and Outlier Analysis of azithromycin prescriptions in January 2009



3 Results

3.1 Time series analysis

The individual antibiotic substances are prescribed differently throughout the year, depending on the application or bacteria causing certain infections. The selection of a suitable antibiotic is at the discretion of the attending physician. This is also reflected in the temporal analysis of the antibiotic prescriptions data for the years 2005 to 2011, which showed that the temporal prescription behaviour, especially the seasonal course, differs according to the individual antibiotic substances. While some substances show a more or less pronounced seasonal pattern (amoxicillin, azithromycin, clarithromycin, doxycycline, roxithromycin), others are prescribed relatively evenly over the year (ciprofloxacin, clindamycin) or have no specific pattern (cefuroxime, phenoxymethyl penicillin, levofloxacin-ofloxacin, sulfamethoxazole-trimethoprim). The clearest seasonal pattern is shown by the antibiotics of the macrolide group (azithromycin, clarithromycin, roxithromycin) with maximum prescriptions in the winter season and minimum values during the summer months.

Regarding the trend component, most antibiotics indicate a decreasing trend. Only amoxicillin, cefuroxime and levofloxacin-ofloxacin reveal an increasing trend, while for azithromycin and clindamycin there is almost no trend apparent.

3.2 Cluster and Outlier analysis

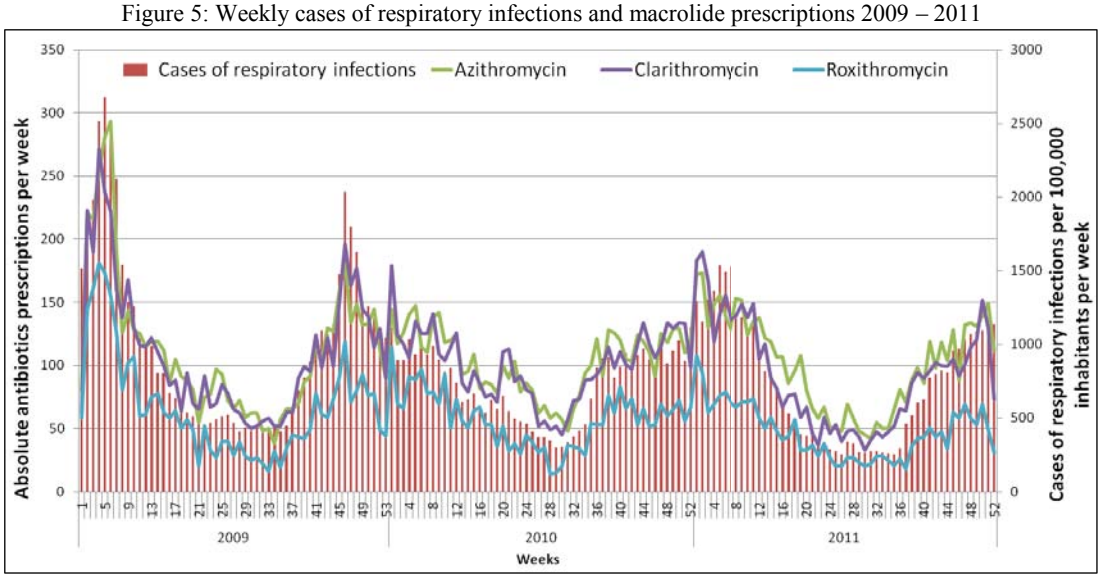
The result of the Cluster and Outlier analysis in ArcGIS depends on the selected distance and spatial relationship of the neighbouring features by which the algorithm calculates the clusters. Therefore, the analysis was first carried out with different parameter settings to find the appropriate preferences.

As conceptualization of spatial relationship *ZONE_OF_INDIFFERENCE* is chosen, where a threshold value specifies whether to include or exclude neighbours and where neighbours are weighted by the Inverse Distance Method after reaching this threshold value. This conceptualization of spatial relationship seemed most suitable for the analysis of the antibiotic prescriptions within the Dresden urban districts, as the polygons of the urban districts have different sizes and the (daily) mobility of citizens may lead to movements across several district boundaries. To cope with these cases it is recommended to use a distance-based conceptualization with a smooth transition as by *ZONE_OF_INDIFFERENCE*.

The ArcGIS tool *Incremental Spatial Autocorrelation*[8] is used for choosing an appropriate threshold distance. The tool calculates the spatial autocorrelation for various distances, to determine the distance at which the spatial processes (the clustering) are most pronounced. This resulted in proper threshold distances of 2800 and 6400 meters for the data at hand, depending on the scale considered for the resulting clusters.

[8] http://resources.arcgis.com/en/help/main/10.1/index.html#//005p0000004z000000

Figure 5: Weekly cases of respiratory infections and macrolide prescriptions 2009 – 2011



Source: respiratory infections data: Google Flu Trends (www.google.org/flutrends/); prescription data: AOK PLUS (AOK Sachsen und Thüringen)

The outcome of these analyses provided resulting cluster maps, which are very different according to the considered antibiotic substance and the given point in time. There were hardly any general patterns for all data. Therefore, the individual cases have to be considered. Nonetheless, there are some urban districts, mostly in the central region of Dresden, that are more often part of a significant cluster than others (see also Figure 4). Thus, from a sewage treatment perspective, these cluster areas are the neighbourhoods that should be given priority and be investigated in more detail.

3.3 Correlation analysis

As stated above data about respiratory infections are considered as a proxy for antibiotic prescriptions. A correlation analysis on the monthly Google Flu Trends data and the prescriptions for all individual antibiotic substances has been performed (Table 1). As result there are three substances presenting a significant strong correlation with the cases of respiratory infections: azithromycin ($r = 0.94$, $p < 0.01$), roxithromycin ($r = 0.83$, $p < 0.01$) and clarithromycin ($r = 0.76$, $p < 0.01$), all belonging to the macrolide group. These three antibiotics show a seasonal pattern similar to the annual wave of influenza.

Consequently the correlation analysis for the three macrolide substances has been repeated with higher resoluted weekly data to include a wider sample and to create a basis for prospective weekly predictions (Figure 5). That way the correlation coefficient has been improved for roxithromycin ($r = 0.88$, $p < 0.01$) and clarithromycin ($r = 0.90$, $p < 0.01$), for azithromycin ($r = 0.91$, $p < 0.01$) it remains almost as high as with the monthly data.

Table 1: Pearson correlation coefficients (R) for monthly Google Flu Trends data and substance-specific antibiotic prescriptions from 2005 to 2011 (n = 84)

Antibiotic	active ingredient groups	R
Azithromycin	Macrolide	0.94**
Roxithromycin	Macrolide	0.83**
Clarithromycin	Macrolide	0.76**
Amoxicillin	β-Lactam	0.61**
Levofloxacin-ofloxacin	Fluoroquinolone	0.53**
Doxycycline	Tetracycline	0.49**
Sulfamethoxazole-trimethoprim	Sulfonamide	0.49**
Phenoxymethyl penicillin	β-Lactam	0.44**
Ciprofloxacin	Fluoroquinolone	0.41**
Cefuroxime	β-Lactam	0.27*
Clindamycin	Lincosamide	0.03

** The correlation is significant at the 0.05 level.*
*** The correlation is significant at the 0.01 level.*

4 Conclusion and Outlook

The presented study succeeded in identifying spatial and temporal patterns of antibiotic descriptions, offering a promising path for future predictions of antibiotic releases in urban waste water. This work also demonstrated a statistical significant correlation between respiratory infections and prescriptions of antibiotic substances of the macrolide group. However, it should be noted that it is not known when exactly a respiratory infection occurs within a week, how long it lasts and when exactly any antibiotics are prescribed or a secondary infection occurs. Moreover, as a recent study pointed to some issues in the usage of Google Flu Trends [14], further

investigation are required to validate the first results, presented in this paper. Additional correlation analyses considering other influencing factors as meteorological and sociodemographic parameters are ongoing.

Future work will focus (1) on the validation of the input model using the results of the ANTI-Resist measurement campaigns and (2) on the development of a prediction model to derive forecasts of the expected antibiotics input into the sewer system. Predictions will be deduced by the combination of the findings of the input model with up-to-date information using regression and interpolation functions. Based on the shown high correlation between respiratory infections and macrolide prescriptions, the weekly available Google Flu Trends data can be integrated into a simple linear regression model to derive prediction intervals of expected antibiotic inputs into the sewer system for the same week. Thus, the final results could be useful for the sewage treatment plant, which could initiate prompt provisions to increasing antibiotics input events, as well as for the public health sector, which could regulate the prescription behaviour of antibiotics in the appropriate way.

Acknowledgements

The project is funded by the Federal Ministry for Education and Research (BMBF) and is part of the program "Research for Sustainable Development". The fruitful cooperation with our ANTI-Resist project partners is gratefully acknowledged. Special thanks go to Daniel Kadner for developing the ANTI-Resist Geoportal.

References

[1] A. L. Batt, S. Kim and D. S. Aga. Comparison of the occurrence of antibiotics in four full-scale wastewater treatment plants with varying designs and operations. Chemosphere, 68(3): 428-435, 2007.

[2] C. L. Cheng, Y. C. Chen, T. M. Liu and Y. H. K. Yang. Using spatial analysis to demonstrate the heterogeneity of the cardiovascular drug-prescribing pattern in Taiwan. BMC public health, 11(1): 380, 2011.

[3] R. B. Cleveland, W. S. Cleveland, J. E. McRae and I. Terpenning. STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1): 3-73, 1990.

[4] P. Davey, C. Pagliari and A. Hayes. The patient's role in the spread and control of bacterial resistance to antibiotics. Clinical Microbiology and Infection, 8(s2), 43-68, 2002.

[5] Forschungsverbund Public Health Sachsen und Sachsen-Anhalt. ANTI-Resist - Online available: http://anti-resist.de/; last accessed 02/2014.

[6] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M.S. Smolinskiand L. Brilliant. Detecting influenza epidemics using search engine query data. Nature, 457(7232): 1012-1014, 2009.

[7] Google, editor. Google Flu Trends. Frequently asked questions, 2011. Online available: http://www.google.org/flutrends/intl/ee_us/about/faq.html; last accessed 02/2014.

[8] H. Goossens, M. Ferech, R. Vander Stichele and M. Elseviers. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. The Lancet, 365(9459): 579-587, 2005.

[9] T. Heberer. Tracking persistent pharmaceutical residues from municipal sewage to drinking water. Journal of Hydrology, 266(3): 175-189, 2002.

[10] S. Herold. Pathogenese, Klinik und Therapie der Virusgrippe. Vom harmlosen Infekt bis zur Intensivstation. Pharmazie in unserer Zeit, 40(2): 115-119, 2011. (In english: Pathogenesis, clinic and therapy of influenza).

[11] W.V. Kern, K. Nink, M. Steib-Bauert and H. Schröder. Regional variation in outpatient antibiotic prescribing in Germany. Infection, 34(5): 269-273, 2006.

[12] K. Kümmerer. Antibiotics in the aquatic environment – a review – part I. Chemosphere, 75(4): 417-434, 2009.

[13] K. Kümmerer. Antibiotics in the aquatic environment – a review – part II. Chemosphere, 75(4): 435-441, 2009.

[14] D. Lazer, R. Kennedy, G. King and A. Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. Science, 343 (6176): 1203-1205, 2014.

[15] F. Modarai, K. Mack, P. Hicks, S. Benoit, S. Park, C. Jones, S. Proescholdbell, A. Ising and L. Paulozzi. Relationship of opioid prescription sales and overdoses, North Carolina. Drug and alcohol dependence, 132(1): 81-86, 2013.

[16] B. Pauwels and W. Verstraete. The treatment of hospital wastewater: an appraisal. J Water Health, 4: 405-416, 2006.

[17] G. Qiu, Y. Song, P. Zeng, L. Duan, and S. Xiao. Combination of upflow anaerobic sludge blanket (UASB) and membrane bioreactor (MBR) for berberine reduction from wastewater and the effects of berberine on bacterial community dynamics. Journal of hazardous materials, 246: 34-43, 2013.

[18] T. Schaber. Diagnostik, Therapie und Prävention der Influenza (Virusgrippe). Pneumologie, 57(01): 27-33, 2003. (In english: Diagnosis, Therapy and Prevention of Influenza).

[19] B. Spellberg, R. Guidos, D. Gilbert, J. Bradley, H. W. Boucher, W. M. Scheld, J. G. Barlett and J. Edwards.

The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. Clinical Infectious Diseases, 46(2): 155-164, 2008.

[20] S. Stone, R. Gonzales, J. Maselli and S. R. Lowenstein. Antibiotic prescribing for patients with colds, upper respiratory tract infections, and bronchitis: a national study of hospital-based emergency departments. Annals of emergency medicine, 36(4): 320-327, 2000.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Influence of point cloud density on the results of automated Object-Based building extraction from ALS data

Ivan Tomljenovic
Department of Geoinformatics (Z_GIS),
University of Salzburg
Schillersrasse 30, 5020, Salzburg, Austria
tomljenoviciv@stud.sbg.ac.at

Adam Rousell
Geographisches Institut,
University of Heidelberg
Berliner Strasse 48, D-69120, Heidelberg, Germany
adam.rousell@geog.uni-heidelberg.de

Abstract

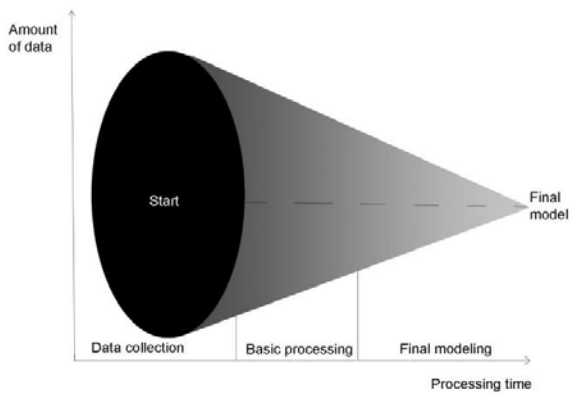
Nowadays there is a plethora of approaches dealing with object extraction from remote sensing data. Airborne Laser scanning (ALS) has become a new method for timely and accurate collection of spatial data in the form of point clouds which can vary in density from less than one point per square meter (ppsm) up to in excess of 200 ppsm. Many algorithms have been developed which provide solutions to object extraction from 3D data sources as ALS point clouds. This paper evaluates the influence of the spatial point density within the point cloud on the obtained results from a pre-developed Object-Based rule set which incorporates formalized knowledge for extraction of 2D building outlines. Analysis is performed with regards to the accuracy and completeness of the resultant extraction dataset. A pre-existing building footprint dataset representing Lake Tahoe (USA) was used for ground truthing. Point cloud datasets with varying densities (18, 16, 9, 7, 5, 2, 1 and 0.5ppsm) where used in the analysis process. Results indicate that using higher density point clouds increases the level of classification accuracy in terms of both completeness and correctness. As the density of points is lowered the accuracy of the results also decreases, although little difference is seen in the interval of 5-16ppsm.

Keywords: ALS, object-based extraction, point density, point clouds, OBIA

1 Introduction

Airborne Laser Scanning (ALS) has become a widely available tool for fast and accurate collection of data. Such a platform is capable of covering large portions of the Earth's surface within short time frames. While having a high speed of collection and thus reducing the necessary time to obtain the data the system is of great benefit, the method also generates a large amount of information. For example, the area of world's smallest country (Vatican 0.2 square miles) could generate records from 500GB up to 2-3TB depending on the chosen point cloud density. New technologies allow the production of great amounts of data in very short time intervals but they still did not provide good solutions for massive point cloud analysis, thus generating a discrepancy between time needed to collect the data and the time needed to process it (figure 1). Because of this, many scientists try to generate faster and more reliable ways of processing the data. Such processes should be as automated as possible thus reducing the influence of the human interpreter in the whole process. One of such approach used in remote sensing is Object-Based Image Analysis (OBIA). [3] recognized that using pixel-based methodologies for data extraction and classification did not provide sufficient results. [2] described in his review the benefits of the Object-Based approach and gave an overview of what has been done in the area so far. By looking at the homogenous units as conceptual wholes one can develop a system of rules which emulate the process of human thinking. By doing this (on a primitive level) it is possible to generate automated processes for data extraction. The process relies on forming the existing knowledge into a set of simplified rules under the framework of Cognition Networking Language (CNL) which is implemented within the eCognition (Trimble) software package.

Figure 1. Graphical depiction of time/data discrepancy when working with ALS data



It must be noted that it became popular to use fused data sources in order to extract information [4, 13, 14, 19, 27, 29, 36], but in our case, we use a single source in order to achieve necessary results. In the work presented here an automated process for building classification based solely on an ALS data source has been developed. For this paper it was decided to test to what extent the usage of different point cloud densities will impact the results obtained from a building extraction classification. The results of the testing will give indications as to if there really is a need to have high density point clouds and how the absence of such a source will influence the outcome. Since the algorithm is converting ALS point clouds into raster images care was also taken with regard the resolution of the data used for the analysis. It was decided to test two approaches. In the first one the resolution is varied based on the point cloud density and in the second one a consistent resolution is used whilst the input density of the point cloud is varied.

2 Previous work on object extraction from ALS data

With the development of ALS technologies and the presence of fast growing spatial data piles, research on the implementation of OBIA methodologies (segmentation and classification) touched fruitful ground. Scientists have developed many approaches which attempt to delineate and classify objects from 3D point clouds with the use of various segmentation based methodologies [1, 5, 6, 9, 10, 12, 16, 17, 18, 19, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35].

When it comes to the generation of the extraction algorithms, most researchers concentrate on domain specific solutions which range from earth surface estimation [8], geomorphic feature detection [7] and Digital Terrain Model creation [21] to modern automatic building extraction [27], automatic road extraction [11, 12], and automatic tree classification [15, 20]. These approaches are producing tangible results but they are not investigating transferability across different ALS data sources. Approaches to point cloud modelling require standardized rule sets which are universally applicable on ALS point clouds. [33] provided one of the earliest descriptions of the extraction process based only on LiDAR data. He used edge detection on an elevation model in order to define candidate objects. A predefined shape assumption (I, T or L shape) was applied in order to extract building objects. [1] used only ALS point cloud data to extract buildings. They used the first minus last pulse method with local statistical interpretation to segment the given data. [25] developed a new method for building extraction in urban areas from high-resolution ALS data. Their approach consisted of DSM minus DTM calculation, height thresholding and the usage of binary morphological operators in order to isolate building candidate regions. [17] provided segmentation and object-based classification methodology for the extraction of building class from ALS DEMs. Their classification was based on regional classification which in turn was based on cluster analysis.

All of previously mentioned approaches utilize point cloud data in order to extract information. This proves that it is possible to obtain tangible information by processing point cloud data. Even though the point clouds are mostly used in order to generate elevation models or raster representations on which the analysis methods are applied, it is still the original ALS data that is being used. Based on these observations and presented use cases an algorithm has been developed for building extraction from ALS data and testing has been performed as to show how point cloud density influences the result of the classification.

3 Methodology

In order to extract tangible objects from ALS data a specific set of rules under the framework of Cognitional Network Language which is a part of eCognition software package were developed. The approach builds on the use of a slope raster generated from the minimum height values of last returns (figure 2b). Based on the slope calculations an initial classification of the scene into hard and weak edges is performed. These classified objects are further refined with the use of pixel growing techniques and based on the analysis of the object's mean height compared to the mean height of the surrounding class it is possible to separate elevated objects from the ground surface.

Figure 2: a) generated Digital Terrain Model (DTM), b) generated digital surface model (DSM) from minimum values of last returns and c) Normalized digital surface model (nDSM) generated by subtracting DTM from DSM

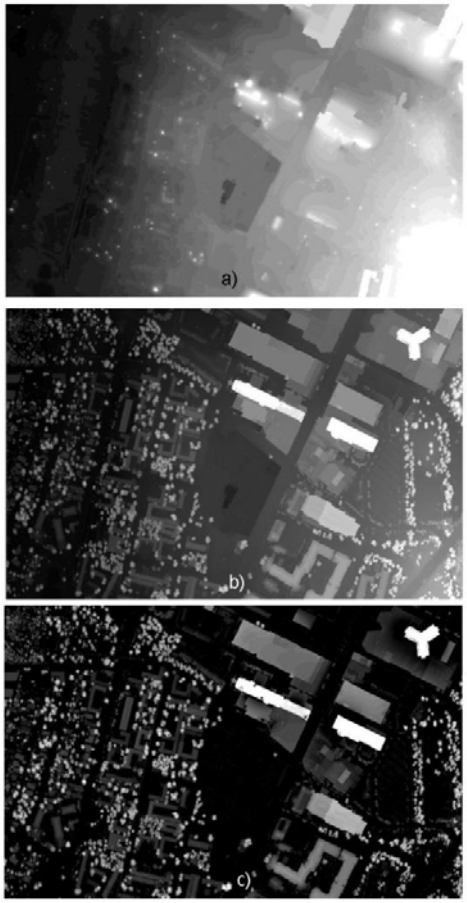
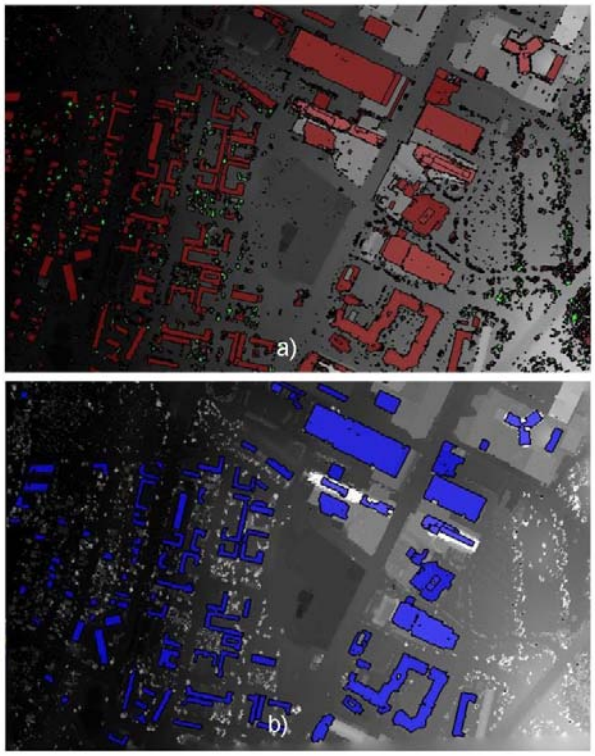


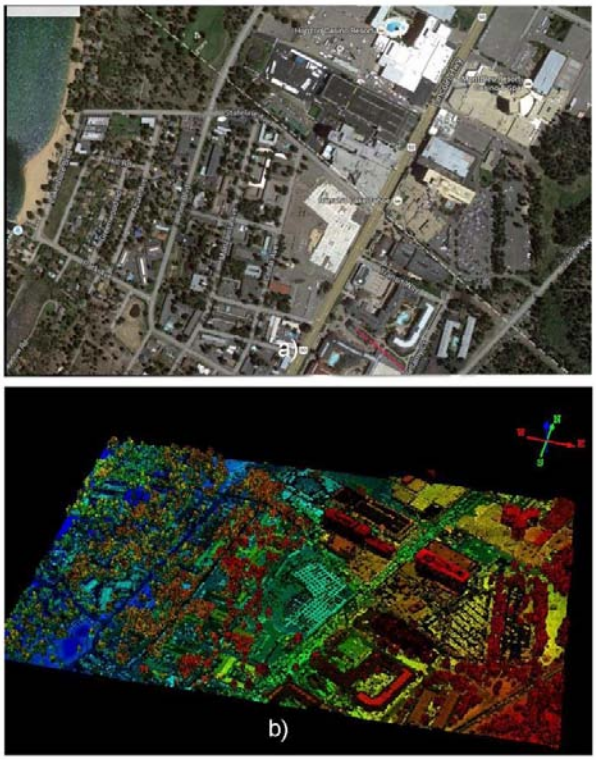
Figure 3: a) Group of objects which represent all the objects which are found above the earth's surface and b) Classified building objects



Based on a number of metrics relating to the objects (intensity, perimeter to area ratio, shape index, rectangular fit and object height) the resulting objects are classified into buildings. The remaining objects are then removed from the classification. Such extracted building objects (Figure 3b) are finally exported to the shapefile format and used for the accuracy assessment. It is important to mention that for the analysis conducted ALS data was used which represents a small area around Lake Tahoe (US) (Figure 4a and 4b). The original point cloud has the density of 18 points per square meter (ppsm).

In order to be able to perform additional testing the Quick Terrain Modeler (Applied Imagery) software was used in order to resample the initial point cloud dataset and generate point clouds with densities of 16, 9, 7, 5, 2, 1 and 0.5ppsm. Each of the newly produced point clouds was then used with the discussed classification algorithm and the resulting objects were exported into shapefile datasets. Since raster representations of surfaces (slope, DTM etc.) were used that were generated from ALS data, it was decided to make two specific use case scenarios. In the first one, the resolution of the rasters were adapted based on the point cloud density (0.25m, 0.5m, 0.75m, 1.0m, 1.5m and finally 2m) and in the second scenario rasters of constant resolution of 0.5 meters were used. All newly generated shape files that were the output of the classification process were compared to the original shape file (reference building data were provided by Spatial Informatics Group operating with funding from the Tahoe Regional Planning Agency) which contains delineated building polygons in order to calculate completes and correctness of our results. The completeness of the classification was calculated by comparing how many objects that were classified as building actually represent buildings. The goal was to compare the classification results for the object and not the absolute correctness of the polygonal shape.

Figure 4: a) Overview of the test data and b) point cloud representing test data



4 Results

Extracted polygons were exported into the shapefile (.shp) format and accuracy measures were performed using the QGIS GIS[1]. Polygon centroids were derived from the extracted polygons and an operation of "point in polygon" was performed to calculate if the extracted polygon is representing a real building polygon by comparing the centroid of the extracted polygon with the building polygons in the reference dataset. In the first case the accuracy measure is generated for the polygons extracted by using resampled point clouds and raster resolution adapted to the point density (table 1). In the second case the accuracy measure is generated for the polygons extracted by using resampled point clouds and raster resolution of 0.5m (table 2).

[1] http:// www.qgis.org/

Table 1: Accuracy results for the first case where the resolution of raster image was adapted to the point cloud density

Shape file – (resolution of raster in meters)	Density (ppsm)	Number of polygons	Polygons representing buildings	Over count	Polygon noise	Completeness (%)	Correctness (overall) (%)	Correctness (from extracted) %
Original Buildings	-	187	187	0	0	100.00%	100.00%	100.00%
Results18-0.25	18	173	154	16	3	90.91%	82.35%	98.27%
Results16-0.50	16	118	95	18	5	60.43%	50.80%	95.76%
Results09-0.50	9	118	91	27	0	63.10%	48.66%	100.00%
Results07-0.50	7	124	90	32	2	65.24%	48.13%	98.39%
Results05-0.75	5	61	38	23	0	32.62%	20.32%	100.00%
Results02-1.00	2	31	14	17	0	16.58%	7.49%	100.00%
Results01-1.50	1	13	6	6	1	6.42%	3.21%	92.31%
Results005-2.00	0.5	7	6	0	1	3.21%	3.21%	85.71%

Table 2: Accuracy results for the second case where the resolution of raster image was constant at 0.5m

Shape file	Density (ppsm)	Number of polygons	Polygons representing buildings	Over count	Polygon noise	Completeness (%)	Correctness (overall) (%)	Correctness (from extracted) %
Original Buildings	-	187	187	0	0	100.00%	100.00%	100.00%
18	18	108	103	5	0	57.75%	55.08%	100.00%
16	16	109	102	7	0	58.29%	54.55%	100.00%
9	9	117	110	7	0	62.57%	58.82%	100.00%
7	7	124	90	34	0	66.31%	48.13%	100.00%
5	5	129	117	12	0	68.98%	62.57%	100.00%
2	2	1	0	0	1	0.00%	0.00%	0.00%
1	1	1	0	0	1	0.00%	0.00%	0.00%
0.5	0.5	1	0	0	1	0.00%	0.00%	0.00%

Table 1 depicts a number of fields. The "Density" column represents the density of the point cloud, "Number of Polygons" represents the total number of polygons extracted with the classification method, "Polygons representing buildings" shows how many of extracted polygons represent a real building polygon, "Overcount" represents polygons which exist due to the over segmentation of a single structure, "Polygon noise" shows misclassified polygons, "Completeness" represents the percentage of extracted true building polygons compared to the actual number of polygons based on the ground truth data, "Correctness (Overall)" shows percentage of extracted polygons which represent single buildings compared to the base data, and "Correctness (from extracted)" represents the percentage of correctly classified extracted polygons within the obtained data. What can be observed from table 1 is that a high level of completeness and correctness (above 80%) was achieved only for the point cloud with the density of over 18ppsm. Point densities between 7-16ppsm show a middle but very stable response and everything below 5ppsm shows a very weak response (under 35%). On the other hand, if the accuracy of the classification from the extracted polygons is observed, it can be noticed that almost all the extracted polygons have above 85% correctness rate.

In the second case (table 2) the level of completeness is stable for the cases from 5-18ppsm and it evolves between the values of 48-63%. Everything below the density of 5ppsm gave a negative response of 0%. If the Correctness of the extracted polygons is observed, it can be noticed that a very high response of 100% for all the cases except the last three densities below 5ppsm is recorded.

5 Discussion & conclusions

Based on the obtained results two observational streams can be identified. In the first case, when the resolution of the data is adapted to the point cloud density, it can be observed that the high point density (18ppsm) along with very high resolution (<0.25m) will provide a high response resulting in increased accuracy. On the other hand, lower point cloud densities (7-16ppsm), along with lower resolution (0.50m), show a stable response when it comes to the accuracy, thus providing the option of using any of these since the resulting outcome will have no significant change in accuracy. In case the resolution is increased further (>0.5m) and decrease the point cloud density (<5ppsm) the results are no longer promising and the algorithm needs to be adapted to the new circumstances (parameter change is required).

In the second case, when using the same resolution of the data and only changing the point cloud densities, it is clear that the obtained response is stable for the point cloud densities of 5ppsm and above, but below 5ppsm the results are

completely deteriorated and thus make a change in the algorithm necessary for such instances.

Based on the obtained results it can be determined that all the point cloud data collected with the point densities of above 5ppsm and with the resolution higher than 0.5m (if rasterisation is applied) can be used with the developed classification approach. In these cases the classification process will provide similar results thus eliminating the need from using more expensive ALS systems which provide very high densities for the collected data. The accuracy of the extraction when it comes to the internal accuracy of extracted objects is very high (>85%) which also shows that the developed algorithm proves the usefulness of OBIA methodologies when applied to 3D data sources which do not mimic human vision. Future work should focus on adapting the existing parameters (shape index, rectangular fit, perimeter to area ratio, number of returns and intensity) in order to increase the extraction accuracy of the polygons from the data so that even higher levels of completeness can be achieved through our automated approach. One of the currently considered approaches is the usage of Agent Based Modelling in order to adapt the parameters automatically based on the input data.

The presented work is framed within the Doctoral College GIScience (DK W 1237N23). The research of this work is funded by the Austrian Science Fund (FWF).

Reference building data were provided by Spatial Informatics Group operating with funding from the Tahoe Regional Planning Agency.

References

[1] A. Alharthy and J. Bethel. Heuristic filtering and 3D feature extraction from LiDAR data. In *Computer Society Conference on Computer Vision and Pattern Recognition*. Kauai, Hawaii, USA. 2001

[2] T. Blaschke. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*. 65(1), pages 2–16. doi:10.1016/j.isprsjprs.2009.06.004. 2010

[3] T. Blaschke and J. Strobl. What 's wrong with pixels? Some recent developments interfacing remote sensing and GIS. *GIS–Zeitschrift Für Geoinformationssysteme*. Pages 12–17. 2001

[4] L. Cheng, L. Tong, Y. Chen, W. Zhang, J. Shan, Y. Liu and M. Li. Integration of LiDAR data and optical multi-view images for 3D reconstruction of building roofs. *Optics and Lasers in Engineering*, 51(4), pages 493–502. doi:10.1016/j.optlaseng.2012.10.010. 2013

[5] W. Cho, Y. Jwa, H. Chang and S. Lee. Pseudo-Grid Based Building Extraction Using Airborne LIDAR Data. In *ISPRS Congress Istanbul 2004*. Istanbul, Turkey. pages 3–6. 2004

[6] F.M.B. Van Coillie, F. Devriendt and R.R. DeWulf. Directional local filtering assisting individual tree analysis in closed forest canopies using VHR optical and LiDAR data. In *Proceedings of the 4th GEOBIA*. Rio de Janeiro. pages 350–354. 2012

[7] P. Dorninger and B. Székely. Automated Detection and Interpretation of Geomorphic Features in LiDAR Point Clouds. *Vermessung & Geoinformation*, (2), pages 60–69. 2011

[8] M. Elmqvist. Ground surface estimation from airborne laser scanner data using active shape models. In *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 34. pages 114–118. 2002

[9] L. Eysn, M. Hollaus, K. Schadauer and N. Pfeifer. Forest Delineation Based on Airborne LIDAR Data. *Remote Sensing*, 4(3), pages 762–783. doi:10.3390/rs4030762. 2012

[10] N. Haala and C. Brenner. Extraction of buildings and trees in urban environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54, pages 130–137. 1999

[11] J. Han, D. Kim, M. Lee and M. Sunwoo. Enhanced Road Boundary and Obstacle Detection Using a Downward-Looking LIDAR Sensor. *In IEEE Transactions on Vehicular Technology* (Vol. 61). pages 971–985. 2012

[12] X. Hu and C.V. Tao. Automatic road extraction from dense urban area by integrated processing of high resolution imagery and LiDAR data. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXV(B3)*. pages 288–292. 2004

[13] D. Li. Remotely sensed images and GIS data fusion for automatic change detection. *International Journal of Image and Data Fusion*, 1(1), pages 99–108. doi:10.1080/19479830903562074. 2010

[14] Y. Li, H. Wu, R. An, H. Xu, Q. He and J. Xu. An improved building boundary extraction algorithm based on fusion of optical imagery and LiDAR data. *Optik - International Journal for Light and Electron Optics*. doi:10.1016/j.ijleo.2013.03.045. 2013

[15] Y. Livny, F. Yan, M. Olson and B. Chen. Automatic Reconstruction of Tree Skeletal Structures from Point Clouds. In *ACM Trans. Graph*. 29, 6, page 151. 2007

[16] L. Matikainen, J. Hyyppa and H. Hyyppa. Automatic detection of buildings from laser scanner data for map updating. In *International Archives of the Photogrammetry and Remote Sensing*, vol. 34, part 3/W13. Dresden, Germany. pages 218–224. 2003

[17] G. Miliaresis and N. Kokkas. Segmentation and object-based classification for the extraction of the building class from LiDAR DEMs. *Computers & Geosciences*, 33(8), pages 1076–1087. doi:10.1016/j.cageo.2006.11.012. 2007

[18] C. Nardinocchi and M. Scaioni. Building extraction from LIDAR data. In *IEEE/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*. Rome, Italy. pages 79–83. 2001

[19] J.P.M. O'Neil-Dunne, S.W. MacFaden, A.R. Royar and K.C. Pelletier. An object-based system for LiDAR data fusion and feature extraction. *Geocarto International*, (August 2012), pages 1–16. doi:10.1080/10106049.2012.689015. 2012

[20] H. Park and R.T. Russeltsfnswgovau. 3D Modelling of Individual Trees Using Full-waveform Lidar. *In Asian Conference on Remote Sensing (ACRS)*. 2009

[21] G.T. Raber, J.R Jensen, S.R. Schill and K. Schuckman. Creation of Digital Terrain Models Using an Adaptive Lidar Vegetation Point Removal Process. *Photogrammetric Engineering and Remote Sensing*, 68(12), pages 1407–1431. 2002

[22] J. Reitberger, C. Schnörr, P. Krzystek and U. Stilla. 3D segmentation of single trees exploiting full waveform LIDAR data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(6), pages 561–574. doi:10.1016/j.isprsjprs.2009.04.002. 2009

[23] F. Rottensteiner. Automatic generation of high-quality building models from LiDAR data. In *IEEE Computer Graphics and Applications*. pages. 42–50. 2003

[24] F. Rottensteiner. Automation of object extraction from LiDAR in urban areas. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Honolulu, Hawaii, USA. pages 5–8. 2010

[25] F. Rottensteiner and C. Briese. A new method for building extraction in urban areas from high-resolution LiDAR data. In *C. Armenakis & Y. C. Lee (Eds.), Commission IV Symposium "Geospatial Theory, Processing and Applications."* Ottawa, Canada. 2001

[26] F. Rottensteiner and J. Jansa. Automatic extraction of buildings from LiDAR data and aerial images. In *Proc. 21st Int'l Soc. Photogrammetry and Remote Sensing Congress (ISPRS)*. Ottawa, Canada. 2002

[27] G. Sohn and I. Dowman. Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(1), pages 43–63. doi:10.1016/j.isprsjprs.2007.01.001. 2007

[28] S. Solberg, E. Naesset and O.M. Bollandsas. Single tree segmentation using airborne laser scanner data in a structurally heterogeneous spruce forest. *Photogrammetric Engineering and Remote Sensing*, 72(12), pages 1369–1378. 2006

[29] A. Swatantran, R. Dubayah, D. Roberts, M. Hofton and J.B. Blair. Mapping biomass and stress in the Sierra Nevada using LiDAR and hyperspectral data fusion. *Remote Sensing of Environment*, 115(11), pages 2917–2930. doi:10.1016/j.rse.2010.08.027.2011

[30] P. Tian and L. Sui. Building contours extraction from light detect and ranging data. In *2011 Symposium on Photonics and Optoelectronics (SOPO)*. Wuhan, China: IEEE. pages 1–3. doi:10.1109/SOPO.2011.5780523. 2011

[31] N.R. Vaughn, L.M. Moskal and E.C. Turnblom. Tree species detection accuracies using discrete point lidar and airborne waveform LiDAR. *Remote Sensing*, 4(2), pages 377–403. doi:10.3390/rs4020377. 2012

[32] G. Vosselman and H.G. Maas. Airborne and Terrestrial Laser Scanning. *Dunbeath: Whittles Publishing*. 2010

[33] Z. Wang and T. Schenk. Extracting building information from LiDAR data. *In ISPRS Proceedings of Commission III Symposium on Object Recognition and Scene Classification from Multispectral and Multisensor Pixels*. Columbus, Ohio. pages 279–284. 1998

[34] Z. Wang and T. Schenk. Building extraction and reconstruction from LiDAR data. *In ISPRS Congress Amsterdam 2000* (Vol. XXXIII). Amsterdam, Netherlands. pages 958-964. 2000

[35] H. Weinacker, B. Koch, U. Heyder and R. Weinacker. Development of filtering , segmentation and modelling modules for lidar and multispectral data as a fundament of an automatic forest inventory system. In *Proceedings of the ISPRS Working Group VIII/2*. Freiburg, Germany. pages 50–55. 2002

[36] J. Zhang. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1(1), pages 5–24. doi:10.1080/19479830903561035. 2010

Session:
Routing

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Street Network created by Proximity Graphs: Its Topological Structure and Travel Efficiency

Toshihiro Osaragi
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo, Japan
osaragi@mei.titech.ac.jp

Yuko Hiraga
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo, Japan
hiraga@os.mei.titech.ac.jp

Abstract

There exists a large body of basic research on street networks using proximity graphs from various viewpoints. In the present study, we employ proximity graphs based on β-skeletons which change in response to variations in parameter values of β, and attempt to analyze street networks from the viewpoint of the topological structure and the travel efficiency at the same time. Some new findings on their relationships are demonstrated by numerical case studies on street networks created by proximity graphs.

keywords: street network; proximity graph; β-skeleton; topological structure; travel efficiency; spanning ratio

1 Introduction

Proximity graphs, also called neighborhood graphs, are simply graphs in which two vertices are connected by an edge if and only if the vertices satisfy particular geometric requirements. "Proximity" here means spatial distance, and many of them can be formulated with respect to many metrics, but the Euclidean metric is used most frequently [4]. These graphs are utilized in multiple applications. For instance, in computer science, properties, bounds on the size, algorithms, and variants of the proximity graphs were discussed, and numerous applications including computational morphology, spatial analysis, pattern classification, and data bases for computer vision were described [7].

In spatial analysis, Tanimura and Furuyama [16] and Watanabe [18] created familiar proximity graphs (Delaunay triangulations, Gabriel graphs, relative neighbourhood graphs, and minimum spanning trees) using the locations of intersection points in actual street networks, and discovered that such networks resemble proximity graphs.

One other area of interest where proximity graphs find application is in the field of transportation, where a graph representation of infrastructure can be used to assess efficiency of travel, configuration, properties of street networks. For instance, Koshizuka and Kobayashi [12] analyzed street networks by looking at the efficiency of travel, specifically, the ratio between shortest path length and Euclidean distance. This ratio is called "*spanning ratio*", which has been studied theoretically and numerically using proximity graphs. Eppstein [6] discussed the dilation of various proximity graphs, defined as the maximum ratio between shortest path length and Euclidean distance. Bose [3] and Wang et al. [17] discussed theoretically the spanning ratio of a proximity graph defined on n points in the Euclidean plane, and obtained the upper-bounds and lower-bounds of the spanning ratio. Watanabe [19] evaluated the configuration and the travel efficiency on proximity graphs.

Thus, proximity graphs have been investigated from two different perspectives. From a morphological perspective the authors mainly focused on topological structure of street networks created by proximity graphs, that is, the ways in which intersections were connected [1, 13, 20]. A different approach that is relevant in transportation is the efficiency of travel, which provides an alternative perspective on networks [6, 12, 19].

In this paper, our objective is to employ the concept of β-skeleton which changes in response to variations in single parameter value of β, in order to investigate street networks from the above two different perspective: the topological structure and the travel efficiency at the same time. The original contribution of this paper is to clarify their relationships which vary according to local geographic characteristics.

2 Topological Structure of Proximity Graphs

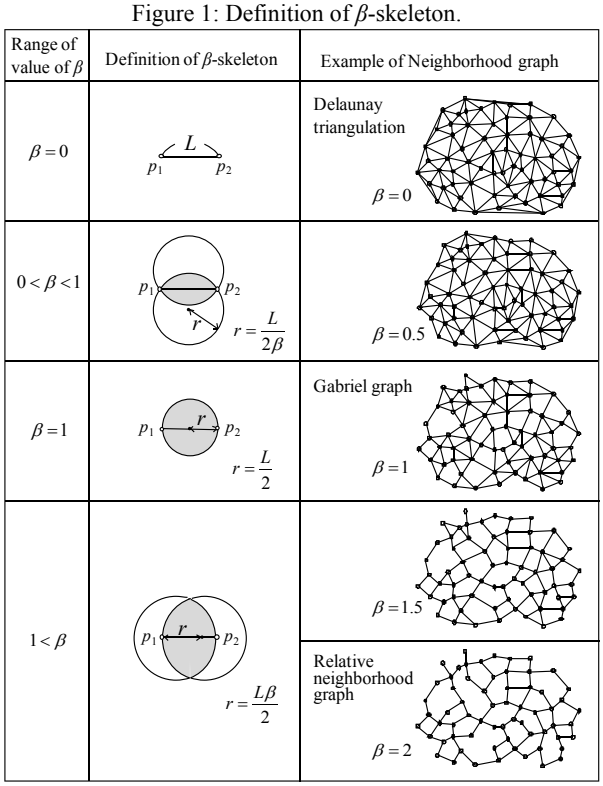
2.1 Concept of β-skeleton

Given a spatial distribution of points p_i (i = 1, 2, …, n) in two-dimensional space, let us consider various ways of creating proximity graphs that connect the points to each other. As shown in Figure 1, let us assume that two circular arcs pass through the arbitrary points p_1 and p_2. The size of the closed region E enclosed by the arcs (the crosshatched portions in Figure 1) varies with the parameter β (≥ 0), such that the area of E increases as β increases. Then, if some third point is included within E, then the segment with endpoints p_1 and p_2 is not an edge in the graph, whereas if no such third point is included, the graph contains this segment as an edge.

A proximity graph created according to this rule is called the β-skeleton and its effective calculation methods were proposed [2, 4, 5, 11, 17]. It is well established that the case β = 0 corresponds to the complete graph (CG), β = 1

corresponds to the Gabriel graph (GG), and $\beta = 2$ corresponds to the relative neighbourhood graph (RNG).

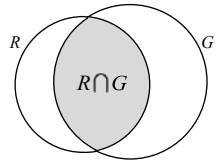
Figure 1: Definition of β-skeleton.



2.2 Definition of agreement rate

Let us define an "agreement rate" as an index expressing how closely the morphology or topology of a proximity graph resembles that of an actual street network (that is, the degree of morphological or topological structure [8, 9]). First, the set of edges making up the street network is denoted by R, and the set of edges making up the proximity graph is denoted by G. The number of elements in the set of edges (number of edges) is written as the function $n(\ )$. Then, we define the agreement rate (C-rate) as the number of elements in $R \cap G$ divided by the number of elements in $R \cup G$, that is, $n(R \cap G)/n(R \cup G)$. Also, we distinguish between what we call the "R-rate", an alternative agreement rate based on the actual street network R, $n(R \cap G)/n(R)$, and the "G-rate", an alternative agreement rate based on the proximity graph G, $n(R \cap G)/n(G)$.

Figure 2: Definitions of agreement rates.



R: Set of road network
G: Set of proximity graph

$$C - rate = \frac{n(R \cap G)}{n(R \cup G)}$$

$$R - rate = \frac{n(R \cap G)}{n(R)}$$

$$G - rate = \frac{n(R \cap G)}{n(G)}$$

2.3 Maximum agreement rate and value of β

A part of the greater Tokyo metropolitan region was chosen as the area for analysis (Figure 3). The analytical region was subdivided into eight subregions according to map borders (as indicated by the numerals in the figure), and each subregion was analyzed in order to consider local characteristics. The highways in each subregion were extracted as the actual street network R (Figure 4). Because the objective is to analyze similarity of topological structure, all the streets between the intersection points of the street network were replaced with straight lines.

Figure 3: Study area.

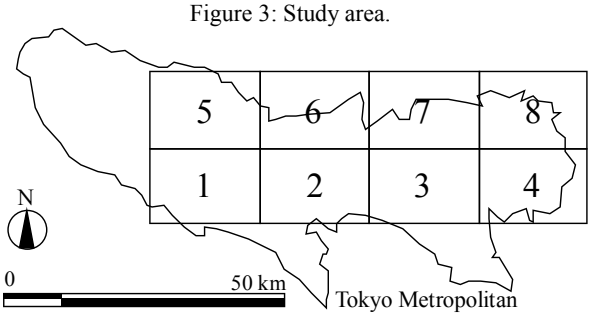


Figure 4: Street networks to be analyzed as R.

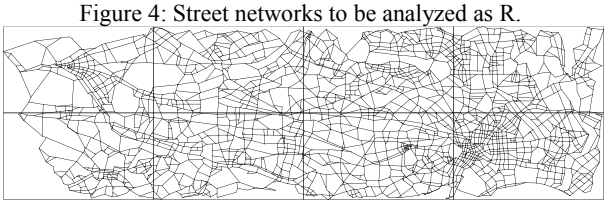


Figure 5 is a set of proximity graphs G in which β is varied from 1.0 to 2.0 in steps of 0.5 using the actual intersection points in Figure 4. As seen, the number of edges decreases gradually as the value of β increases.

In Figure 6 (a), the edges in the actual street network R are shown with the portion common with the proximity graph G ($\beta = 1.5$) ($R \cap G$) indicated by thick lines. In Figure 6 (b), the proximity graph G ($\beta = 1.5$) is shown, again with the common portion with the actual street networks ($R \cap G$) indicated by thick lines.

Proximity graphs G were created for various values of β, using Subregion 4 as an example, and the resulting C-rate, R-rate, and G-rate with respect to the actual street network were calculated (shown in Figure 7). The value of β yielding the maximum agreement rate is labeled β_1. There is a trade-off between maximizing the G-rate and maximizing the R-rate, but the agreement rate (C-rate) is a comprehensive index providing a balance between the two.

The agreement rate (C-rate) for each of the eight subregions in the study area were calculated after creating proximity graphs G for various values of β. Table 1 shows the maximum agreement rate and the corresponding β_1. As shown, the values of β_1 for the subregions lie between 1.1 and 1.5.

Figure 5: Proximity graphs based on β-skeletons for different values of β.

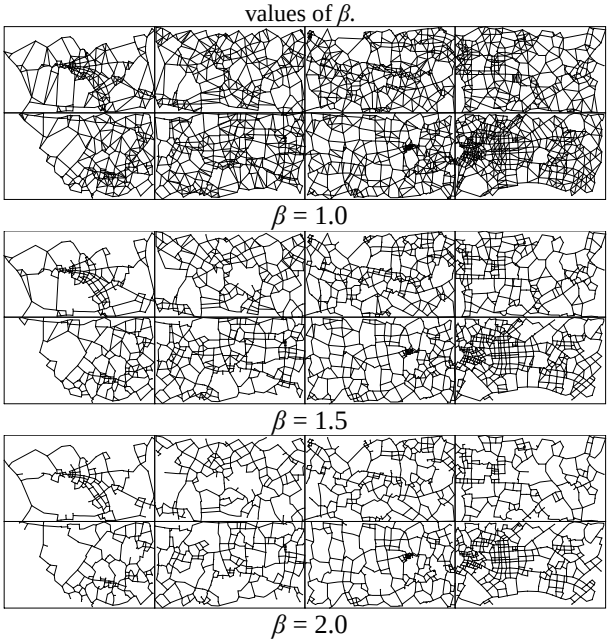


Figure 6: Common (thick lines) and disjoint (thin lines) edges of the street network R (a) and proximity graph G (b), where β = 1.5 for Subregion 4.

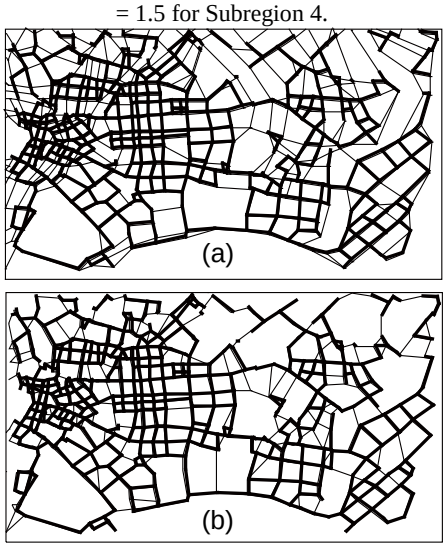


Figure 7: Agreement rate as function of β for Subregion 4.

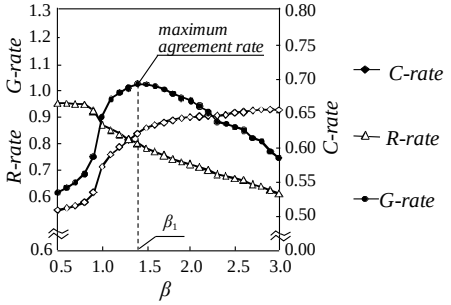


Table 1: Maximum agreement rate and the corresponding value of β_1.

Subregion	Maximum agreement rate	β_1
1	0.610	1.40
2	0.643	1.45
3	0.639	1.15
4	0.693	1.40
5	0.623	1.20
6	0.614	1.20
7	0.637	1.30
8	0.656	1.25

2.4 Relation between maximum agreement rate and density of intersection points

Figure 8 demonstrates how the maximum agreement rate varied with the density of intersection points (the number of intersections per square kilometer). The highest β_1 in the Tokyo region is for Subregion 4, where the density of intersection points is greatest; β_1 is lowest in Subregions 1, 5, and 6, which have the low densities of intersection points.

Let us consider why the agreement rate is low for these areas, such as mountainous areas, where the density of intersection points is low. As shown in Figure 9, builders of actual street networks tend to skirt mountainous areas, so spatially neighboring points p_1 and p_2, as well as points p_3 and p_4, are not directory connected to each other. However, in proximity graph G, only the spatial relationships are considered, and so the agreement rate was lowered by the addition of edges between such points.

Figure 8: Maximum agreement rate versus density of intersection points (numerals indicate subregion)

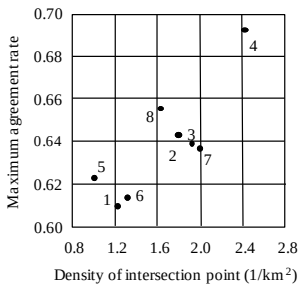
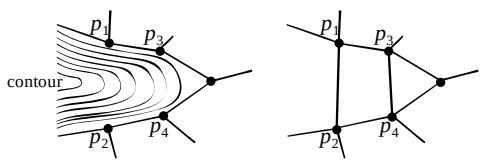


Figure 9: Explanation of low agreement rates for mountainous areas

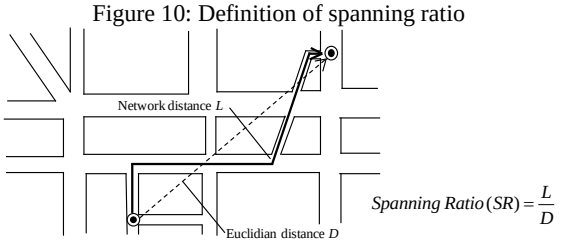


3 Travel Efficiency of Proximity Graphs

3.1 Concept of spanning ratio

The spanning ratio (*SR*) has been suggested as an index expressing the travel efficiency through a network [3, 17]. *SR* is defined as the value of the distance *L* between two points on the network paths divided by the Euclidian distance *D* between the points (Figure 10). In other words, the greater the values *SR*, the lower the travel efficiency in the network.

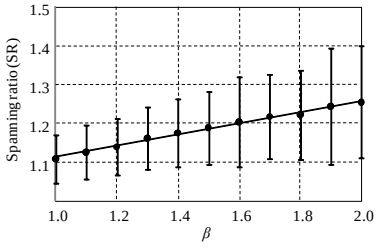
Figure 10: Definition of spanning ratio



3.2 Spanning ratio of proximity graphs

The intersection points in the street networks *R* in the previous section were used to create proximity graphs for various values of β ($1.0 \leq \beta \leq 2.0$). Next, two intersections at a time were extracted at random and the value of *SR* was calculated for that pair. The mean *m* and standard deviation σ were calculated for the *SR* of 1,000 point pairs for each graph. The results showed that *m* is an increasing linear function of β ($m = a\beta + b$; *a* and *b* are unknown parameters). The increase in *m* is due to proximity graphs with higher values of β having lower numbers of edges, decreasing the efficiency of spatial movement in the graphs (Figure 11).

Figure 11: Mean and standard deviation (indicated by error bars) of spanning ratio of proximity graphs G for Subregion 4.



Also, the results showed that the value of σ grows with the value of β. The growth of σ indicates that there is high variation in the travel efficiency between point pairs, that is, that there is a large difference between the Euclidian distance and the network distance between point pairs. Therefore, when we conduct analysis of spatial movement in regions with low street densities, it is preferable to use network distance rather than Euclidian distance.

The mean *m* of *SR* for 1,000 point pairs was calculated for the actual street network of each subregion. The values of β (β_2) were then inversely estimated using *m* by the equations ($\beta_2 = (m - b)/a$). Specifically, the values of β for the proximity graph indicating the mean values of *SR* equivalent to that of the actual street network were calculated. These values are shown in Table 2 along with the corresponding values for parameters of regression equations. As shown, in all the subregions analyzed here, β_2 remains within the range 1.1 to 1.5, the same as β_1.

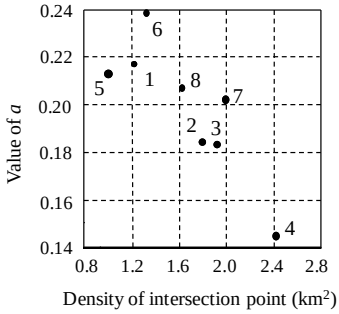
Table 2: Value of β_2 for the proximity graph whose travel efficiency is equivalent to that of actual street network.

Subregion	*m*	*a*	*b*	R^2	β_2
1	1.224	0.217	0.913	0.993	1.440
2	1.196	0.184	0.934	0.993	1.432
3	1.155	0.184	0.946	0.981	1.146
4	1.166	0.145	0.968	0.993	1.363
5	1.184	0.213	0.906	0.998	1.310
6	1.194	0.238	0.874	0.994	1.350
7	1.178	0.202	0.914	0.989	1.310
8	1.210	0.207	0.918	0.995	1.374

3.3 Relation between spanning ratio and density of intersection points

Figure 12 shows how the slopes *a* in Table 2 varied by the density of intersection points. As shown, the lower the density, the greater the slope. Since slope *a* indicates the rate of increase in *SR* with respect to an increase in β (from the regression equation $SR = a\beta + b$), the lower the density of streets in a region, the greater the influence of β on travel efficiency (*SR*) in the corresponding proximity graphs. Thus, the travel efficiency in an area with a low density of intersection points will be more strongly influenced by street closures, for example due to earthquakes, than higher density areas.

Figure 12: Slope *a* versus density of intersection points (numerals indicate subregion)

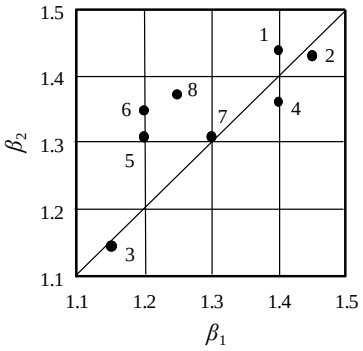


3.4 Relation between β_1 and β_2

Figure 13 shows relationships between the β_1 (value of β for proximity of topological structure) and the β_2 (value of β for proximity of travel efficiency). The values of β_1 and β_2 are roughly similar in subregions 1, 2, 3, 4, and 7, mainly the downtown Tokyo area, where the density of streets is high. On the other hand, in subregions 5, 6, and 8, suburban areas with low densities of streets or areas with mountains or wide rivers, $\beta_1 < \beta_2$ holds. In these areas, there is a risk that using proximity graphs, which have been created on the basis of proximity of topological structure, will provide erroneous predictions of travel efficiency. Specifically, the travel

efficiency in the actual street network is likely to be lower than that in the proximity graph created on the basis of topological proximity in these areas.

Figure 13: β_1 versus β_2 (numerals are subregion numbers).



4 Summary and Conclusions

We carried out an analysis of a street network created by proximity graphs based on β-skeletons from each of two viewpoints, topological structure and travel efficiency. The following findings were identified:

(1) The value of β in a proximity graph with a maximal topological proximity to an actual street network is in the range 1.1 to 1.5 for the networks examined here.

(2) The agreement rate between a street network and a proximity graph is less in mountainous suburban areas or similar areas with low densities of streets.

(3) The value of β in a proximity graph in which travel efficiency is equivalent to an actual street network is in the range 1.1 to 1.5 for the networks examined here.

(4) The travel efficiency (Spanning Ratio: *SR*) between two points shows more variation in suburban areas with low densities of streets; therefore, when investigating the travel efficiency between locations, the analysis must employ the distance in the network rather than the Euclidian distance between the points.

(5) The value of β_1 when there is high topological proximity was nearly equal to the value of β_2 when there is a strong similarity between the travel efficiencies in the central part of Tokyo. However, $\beta_1 < \beta_2$ in the Tokyo suburbs, indicating that an analyst must take account of the higher travel efficiency in the proximity graph mostly strongly resembling the actual street network than that in the actual street network itself.

In this paper, we investigated the properties of proximity graphs by comparing with actual street networks. This approach can be extended for the general modeling of various numerical simulations, as well as theoretical analysis on intersections which are randomly distributed following the Poisson distribution. It would be also interesting to develop this approach for the street hierarchies from the multiple perspectives of topology and geometry [10], and for a method to automate street networks in urban area [14].

Acknowledgements

The authors would like to acknowledge the valuable comments and useful suggestions from Prof. Daisuke Watanabe (Tokyo University of Marine Science and Technology) and anonymous reviewers to improve the content and clarity of the paper.

References

[1] M. Barthélemy and A. Flammini. Modeling urban street patterns, *PHYSICAL REVIEW LETTERS*, 100(13), 138702-1-4, 2008.

[2] M. Bhardwaj, S. Misra and G. Xue. Distributed topology control in wireless ad hoc networks using ß-skeleton, *Workshop on High Performance Switching and Routing* (*HPSR* 2005), Hong Kong, China, 2005.

[3] P. Bose, L. Devroye, W. Evans and D. Kirkpatrick. On the spanning rate of Gabriel graphs and β-skeletons, *Lecture Notes in Computer Science*, 2286, 479-493, 2002.

[4] P. Bose, J. Cardinal, S. Collette, E. D. Demaine, B. Palop, P. Taslakian and N. Zeh. Relaxed Gabriel graphs, *Proc. 15th Canadian Conference on Computational Geometry* (*CCCG* 2009, *Vancouver*), 169-172, 2009.

[5] J. Cardinal, S. Collette and S. Langerman. Empty region graphs, *Computational Geometry*, 42(3), 183-195, 2009.

[6] D. Eppstein. Beta-skeletons have unbounded dilation, *Computational Geometry Theory & Applications*, 23(1), 43–52, 2002.

[7] W. J. Jaromczyk and G. T. Toussaint. Relative neighbourhood graphs and their relatives, *Proc. IEEE*, 80, 1502-1517, 1992.

[8] B. Jiang and C. Claramunt. Topological analysis of urban street networks, *Environment and Planning B*: *Planning and Design*, 31, 151-162, 2004a.

[9] B. Jiang and C. Claramunt. A structural approach to the model generalization of an urban street network, *GeoInformatica*, 8(2), 157-171, 2004b.

[10] B. Jiang. Street hierarchies: a minority of streets account for a majority of traffic flow, *International Journal of Geographical Information Science*, 23(8), 1033-1048, 2009.

[11] F. Hurtado, G. Liotta and H. Meijer. Optimal and suboptimal robust algorithms for proximity graphs, *Computational Geometry Theory & Applications*, 25(1–2), 35–49, 2003.

[12] T. Koshizuka and J. Kobayashi. On the relation between street distance and Euclidean distance, *City planning review*, 18, 43-48, 1983.

[13] S. Porta, P. Crucitti and V. Latora. The network analysis of urban streets: A dual approach, *Physica A*: *Statistical Mechanics and its Applications*, 369(2), 853-866, 2006.

[14] J. Radke and A. Flodmark. The use of spatial decompositions for constructing street centerlines, *Geographic Information Sciences*, 5(1), 15–23, 1999).

[15] C. Ratti. Urban texture and space syntax: some inconsistencies, *Environment and Planning B*: *Planning and Design*, 31, 151-162, 2004.

[16] T. Tanimura and M. Furuyama. A study on the rational network morphology embepped in English historic town, *Journal of architecture, planning and environmental engineering, Transactions of AIJ*, 563:179-186, 2002.

[17] W. Wang, X. Y. Li, K. Moaveninejad, Y. Wang and W. Z. Song. The spanning ratio of β-skeletons, *Proc. 15th Canadian Conference on Computational Geometry* (*CCCG* 2003, *Halifax*), 35–38, 2003.

[18] D. Watanabe. A study on analysing the street network pattern using proximity graphs, *Journal of the City Planning Institute of Japan*, 40, 133-138, 2005.

[19] D. Watanabe. Evaluating the configuration and the travel efficiency on proximity graphs as transportation networks, *FORMA*, 23(2), 81-87, 2008.

[20] D. Watanabe. A study on analyzing the grid road network patterns using relative neighborhood graph, *The Ninth International Symposium on Operations Research and Its Applications* (ISORA'10), 112–119, 2010.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

The effects of different verbal route instructions on spatial orientation

Rui Li, Stefan Fuest, Angela Schwering
Institute for Geoinformatics, University of Muenster
Heisenbergstr. 2, Muenster, Germany
{rui.li; stefan.fuest; schwering}@uni-muenster.de

Abstract

Providing cognitively effective wayfinding instructions is an ongoing research agenda. In addition to providing instructions that are easy to follow, work has started to address instructions that can potentially facilitate spatial orientation and cognitive mapping. In this study, we use a type of verbal instructions that consists of not only landmarks at decision points but also additional landmarks along a route or in distance that are considered crucial for maintaining spatial orientation. The orientation-based route instructions are compared with machine-generated as well as skeletal instructions. Eleven participants were randomly assigned to use one of these three types of instructions to mentally walk a route that they are unfamiliar with and then performed a set of tasks. Preliminary results show that participants using the orientation instructions made fewest errors in their performance of direction estimation. Results from their drawn sketch maps also show more accuracy in global and local orientation. This type of instructions, not surprisingly, does not contribute to accurate estimation of distance. The machine-generated instructions which include distance information, however, are not found contributing to the best estimation of distance. This study supports the potentials of designing wayfinding instructions to facilitate spatial cognition. It also calls the necessity for more comprehensive studies on the effects of instructions on various aspects of wayfinding behaviors, as well as on the automatic generation of orientation-based instructions.

Keywords: Route instructions, landmarks, orientation information, wayfinding, spatial orientation.

1 Introduction

Landmarks are suggested to be crucial in wayfinding instructions to support effective and easy wayfinding as they are indicators of locations in a large-scale environment [1-3]. They have been frequently referred to as decision-making points for reorientation [4]. Studies have shown that constructing wayfinding instructions with local landmarks at decision points lead to more efficient wayfinding [5]. Additionally, research has addressed that landmarks are not only crucial at decision-making points for reorientation [6] but also important along routes for maintaining orientation [7]. We emphasize that cognitively efficient wayfinding instructions should support not only the ease of wayfinding but also spatial orientation during wayfinding. This present study contributes to the understanding of the effects of verbal route instructions including landmarks not only at decision points but also along the route and in distance on spatial orientation and cognitive mapping. In particular, we compare verbal wayfinding instructions including: machine-generated instructions (i.e., Google Maps[1]), our designed instructions with landmarks at decision points, along the route, and in distance (orientation-based instructions), and skeletal instructions with landmarks only at decision points [4].

2 Related work

2.1 Role of landmarks

One important role of landmarks is the identification of particular locations [1] as they are discrete objects or scenes against a background that support the easy identification of locations [2]. Another important role of landmarks is their support for reorientation in wayfinding [4]. Studies have suggested the use of landmarks as a primary or complementary source in wayfinding instructions [3, 8] as they are effective for better outcome such as easier wayfinding guide, fewer wayfinding errors, and shorter wayfinding time [9]. For example, researchers like Tom and Denis [5] compared the use of landmarks in wayfinding instructions with the use of street names. They suggested that using landmarks in wayfinding instructions leads to shorter wayfinding time. Additionally Ross and collaborators [10] found in their study that using landmarks in route instructions leads to less wayfinding errors. In short, the potential of using landmarks in wayfinding instructions is well recognized.

Spatial orientation is also mostly commonly supported by landmarks. As one of important spatial skills, spatial orientation enables persons to be aware of their current locations in relation to destination or other locations in an environment [11]. Wayfinders estimate their locations and relationships between current and other locations in the environment to stay spatially oriented through the use of reference systems [12]. The reference systems could either be egocentric or geocentric [13]. The use of egocentric reference system involves using wayfinders' velocity and acceleration information about their own movement [14], which is less common. In contrast, the use of geocentric reference systems involves the information from the environment. Wayfinders can relate to the features of an environment (i.e. landmarks) and determine the relative locations of themselves or a feature to other features in the environment.

2.2 Location of landmarks

The location of landmarks described in route instructions has intrigued different suggestions in the literature. For example,

[1] http://maps.google.com

Denis and collaborators [15, 16] suggest that wayfinders often use landmarks for reorientation at decision points where a change of direction is necessary. Therefore, no landmark at decision points would become more difficult for wayfinders to determine the locations where they should change heading directions. Moreover, Lovelace and collaborators [6] suggest that landmarks are not only important at locations where reorientation is needed but also essential at locations where change of direction can be possible (potential decision points). At these potential decision points, wayfinders need to maintain their orientation by continuing the same heading direction. They emphasize that having short wayfinding instructions does not automatically translate into good instructions. Consequently, for achieving short and good verbal instructions, Raubal and Winter [3] suggested the use of local landmarks in wayfinding instructions by providing measures to identify the salience of landmarks in an environment. These measures derive from aspects such as visual salience (e.g., facade, shape, color, and visibility), structural salience (e.g., nodes, boundaries, and regions) and semantic salience (e.g., cultural and historic importance of object). Moreover, Richter and Klippel [17] address that the route direction should also be context specific as the structure of the environment is a factor that influences the way how wayfinding instructions should be given. Also aiming to achieve cognitively efficient wayfinding instructions, we introduce a different perspective by looking at the roles of landmarks in instructions that are not only at potential decision points but also along the route as well as in distance.

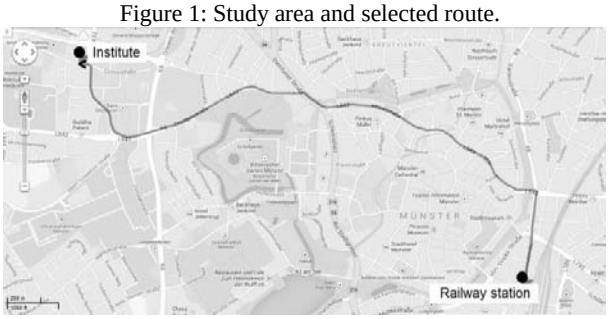
Most of the existing studies introduced above focus on the roles and use of local landmarks that are at potential decision points. Limited studies have addressed the roles of landmarks that are distant from a described route (global landmarks) as those landmarks in distance serves the important role of providing general orientation [18]. Steck and Mallot [19] suggested that one or a couple spatial features could be introduced as global landmarks in wayfinding instructions to provide an initial global orientation. Those global landmarks later could be reintroduced as local landmarks if they are on a designed route [20]. Based on this suggestion, hierarchical communication of space could be achieved by firstly introducing a prominent global feature in instructions, and then specific instructions to maintain orientation and reach destination. In short, the important role of global landmarks has already been remarked. In this paper, we address the use and the role of global landmarks in verbal wayfinding instructions.

In summary, studies have focused on local landmarks and global landmarks in wayfinding. But research on the role of both local and global landmarks for orientation is rather limited. The global landmarks is used adapting the hierarchy suggested by Steck and Mallot [19]. More so, the study of local landmarks was mainly addressing those located at actual or potential decision points. In this paper, we address the use of both local and global landmarks in verbal route instructions. Particularly the location of local landmarks is not only at potential decision points but also along the route. This type of instructions is compared with machine-generated and skeletal instructions as used in previous studies (see [4, 21]) to reveal the different effects on performance of spatial orientation and cognitive mapping.

3 Methods

To construct wayfinding instructions of each type, we selected a route within the city where the university is located. The origin is the central railway station and the destination is our institute building. The length of the selected route is approximately 3.9 km (3 km air distance). The study area and the route from the origin to the destination are shown in Figure 1 below.

Figure 1: Study area and selected route.



Source: Google Maps.

Our primary research goal is to investigate the effects of different route descriptions on the performance spatial orientation without the influence of a person's familiarity with the environment. Therefore, we changed the names of all spatial entities in our verbal descriptions to avoid participants' familiarity. We introduced the study area as a mid-size German city with an old town in its center and a ring-like arrangement of streets. The route itself remained the same shape as in the original route, while the names of street and other spatial entities were changed in instructions. For example, at the original location, the name of railway station was replaced by the name of a fictional cinema, while at the destination the name of institute building was replaced by the name of a fictional library.

Table 1. Three types of wayfinding instructions for the same route segment used in this study.

Type	Instructions
1.Machine-generated	Turn left onto Bismarck street and drive 350m; Continue onto Schiller street for 650m; Continue onto Kreuz street for 140m.
2.Orientation-based	Follow the street, which is heading away from the city center; You cross the intersection on the ring road that runs around the city; Right after you pass the university main building on your right hand side, you reach an intersection.
3.Skeletal	Walk along the street; Right after you passed the university main building, which is on the right side, you reach an intersection.

Three different types of wayfinding instructions have been constructed. Table 1 provides an example of these three types. The first type consists of machine-generated route instructions

from Google Maps. The second type (orientation-based instructions) provides a route description with landmarks not only at potential decision points based on our previous finding [7]. This type of instructions consists of local landmarks at potential decision points and alongside the route, as well as global landmarks in distance. The third type is constructed according to the skeletal descriptions designed by Denis [16] and used in their later studies [5]. This type of instruction consists of a minimum set of wayfinding instructions with landmarks only at decision points.

3.1 Participants

The study was carried out as a pilot. Eleven participants (Age: M = 35.09, SD = 14.35; 7 men and 4 women) were recruited. Participants were not exclusively students.

3.2 Procedure

Participants randomly received one type of wayfinding instructions. They were then asked to complete a set of tasks using the wayfinding instruction they received. The first task of the experiment was drawing a sketch map of the described route from the origin to the destination. In the second task, participants were asked to estimate directions and distances at various locations. This task included three subtasks. The first subtask was estimating the direction back from the destination to the origin of the route (facing the same direction) as well as judging the corresponding air distance. In the second and the third subtask, participants needed to mentally change their position to specific landmarks or intersections (depending on the type of wayfinding instructions) on the route and point to the origin and the destination, and then estimate the air distance in between.

To complete the experiment, each participant was asked to fill in two self-rated measures and one spatial ability test including the Santa Barbara sense of direction scale [22], the spatial anxiety scale [23] and the Purdue spatial visualization test for rotations [24].

4 Results

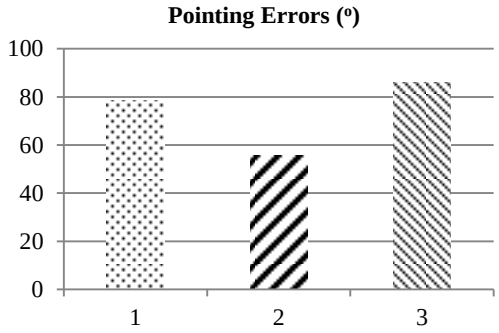
We present the results of our study as follows: 1) direction estimation based on different instructions; 2) distance estimation based on different instructions; 3) sketch maps with respect to route orientation; and 4) the self-rated measures and spatial skills.

4.1 Estimation of direction

Figure 2 shows the average pointing errors among all groups. Participants using the orientation-based instructions made fewest errors in their estimation of direction (M = 55.50°). For the other two instruction groups, the average pointing error are much larger (machine-generated instruction group: M = 78.67°; skeletal instructions group: M = 85.96°). The orientation-based instructions are the only type that includes the city center as a global landmark. Additionally local landmarks are provided not only at decision points but also along the route. Therefore it seems easier, comparing with the other two types of wayfinding instructions, for participants to mentally arrange the described route into a spatial configuration. These landmarks (both global and local ones) included in the instructions facilitates the estimation of directions that requires spatial orientation.

As skeletal instructions consist of the least information, the corresponding construction of mental representation seems very limited. This might also be affected by the lack of landmarks, which seems the same regarding the machine-generated instructions. Locations in the environment could not be unambiguously determined based on these types of instructions.

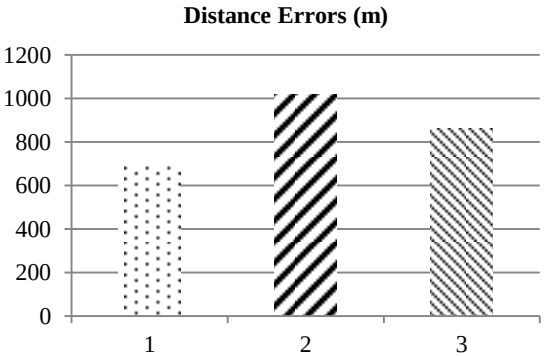
Figure 2. Average pointing errors made by participants among three groups.



4.2 Estimation of distance

As we expected, the distance errors for the machine-generated instructions group are the fewest among all three groups (M = 691.25m). These instructions (see Table 1 for example) include distance information for each route segment. However, what we found surprising is that the machine-generated descriptions did not support so accurate estimation, as the average distance errors are still large. Figure 3 shows the average distance errors among all three groups.

Figure 3. Average estimated distance errors made by participants among all three groups.



The distance errors from the skeletal instructions group have been very large (M = 862.50m). Interestingly, for participants using this type of instructions, the distances that they estimated were distinctively shorter than those in the other two groups. This is likely due to the limited information provided in the skeletal instructions.

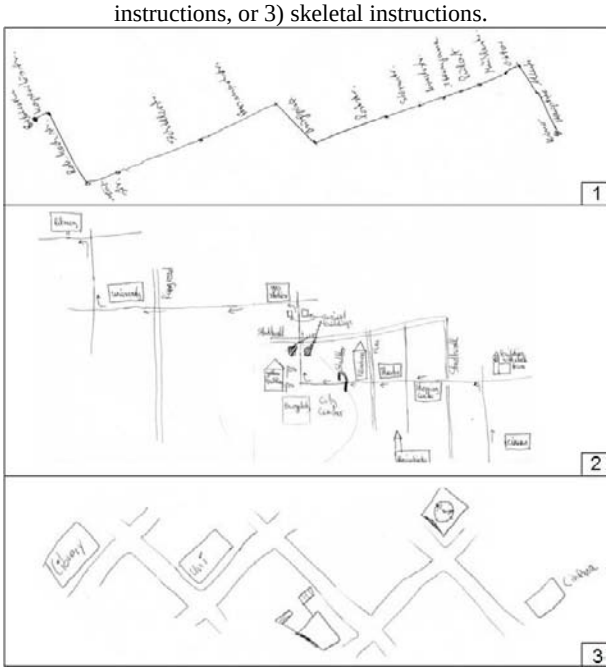
The greatest error was made by participants using the orientation-based instructions (M = 1016.67m). It is not surprising as the instructions do not include distance information. What the results confirm is that both global and

local landmarks provided in instructions do not support the acquisition of the specific metric spatial knowledge: distance.

4.3 Sketch maps

We further analyzed the sketch maps drawn by participants in terms of the orientation of route segments. The original route was divided into major segments at decision points where a big change of direction has occurred. In total we created four major route segments. The same procedure has been used for all sketch maps by identifying the corresponding nodes in the sketch maps. Consequently we measured the angles between these segments and compared them with those on the original route. What we noticed is that the orientation of route segments approximately matches the pattern of participants' direction estimation errors. Particularly, the mean angular error in sketch maps from participants using orientated-based instructions is the fewest (M = 8.56°) among all three types. However the mean angular errors for the other two types of instructions are much greater (machine-generated instructions group: M = 17.08°; skeletal instructions group: M = 13.58°). Figure 4 shows the example of a typical sketched map from each group.

Figure 4. Sample sketch maps drawn by participants using 1) machine-generated instructions, 2) orientation-based instructions, or 3) skeletal instructions.



We also measured the length of each route segment within each sketch maps. Unlike the actual distances of the route segments that vary, participants using skeletal instructions drew each route segment with a very similar length (map 3 in Figure 4). This is primarily due to the very limited information given in this type of instruction that participants were unable to derive distance from the instructions.

In the orientation-based instructions group, the lengths of drawn route segments are more accurate than those in the skeletal instructions group (map 2 in Figure 4). Participants used the described landmarks as references in drawing. It is important to note that sketch maps from this group show different lengths for route segments, which are more accurate than those from the skeletal instructions group. Participants using orientation-based instructions, however, made the greatest error in distance estimation.

In the machine-generated instructions group, the lengths of route segments are relatively more accurate than both other groups (map 1 in Figure 4). As the instructions provide distance information for each segment, participants are likely to draw sketch maps based on this information. This also explains the linear appearance in sketch maps, as well as the fewest errors in their distance estimation task.

4.4 Self-rated measure and spatial skills

The average score of the sense of direction scale (SOD) does not show significant differences among all three groups: 5.20 for participants using machine-generated instructions, 4.42 for participants using orientation-based information, and 4.38 for participants using skeletal descriptions. Regarding the scale of spatial anxiety, participants in the orientation-based instruction group had the highest level of spatial anxiety (4.42), whereas participants in machine-generated instruction group and skeletal description groups have slightly lower spatial anxiety (2.67 and 3.90, respectively). It is interesting to note that participants rated their spatial anxiety the highest in the orientation-based instruction group, but their performance in tasks was not the worst among all three groups. Whether this type of wayfinding instructions can support those who have great spatial anxiety will be further addressed in our ongoing studies.

The score of mental rotation test shows that the participants generally had similar spatial abilities (4.5 for machine-generated instruction group; 5 for orientation-based instruction group; and 4.75 for skeletal description group). Here we only present the descriptive statistics of participants' scores to indicate that participants do not represent great differences among groups. With the involvement of more participants in our continuing study, we intend to investigate the association between these measures and participants' performance using different types of wayfinding instructions.

5 Discussion

5.1 The effects on spatial orientation

As the machine-generated instructions only include distances and street names, it is not surprising that the distance estimation is more accurate than the direction estimation. The biggest challenge for this type of instructions is the acquisition of spatial configuration, as it is not supported by the turn-by-turn instructions. For the skeletal instructions group, it is also very apparent that both distance and direction estimation tasks are difficult, as very limited information is provided. Furthermore, little information with landmarks only at decision points seems to imply short distance for each route segment. This type of route instructions may efficiently guide a person from the start point to the destination, but may not greatly contribute to the person's spatial orientation. For the orientation-based instructions, however, persons are provided with additional landmarks along and distant to the route.

These described landmarks provide confirmation information for guiding a person to reach the destination. Furthermore our preliminary results show the potential of using landmarks that are along a route (local) or in distance (global) to support spatial orientation with directions. Yet this does not lead to accurate estimation of distance.

5.2 Sketch maps

Participants using the machine-generated wayfinding instructions drew sketched maps with very few spatial features. Route segments are most drawn as straight lines. This is primarily caused by the turn-by-turn characteristics of machine-generated instructions. As intersections are not described in this type of instructions, not surprisingly, these sketched maps do not include any spatial entities except streets. Sketched maps based on orientation-based instructions show a spatial configuration of the area in addition to the route described. Additional street segments and more accurate placement of local and global landmarks are also included in sketch maps of this type. It seems that described global landmarks and local landmarks facilitate the acquisition of spatial configuration. Sketch maps based on the skeletal instructions are quite different. Because the wayfinding instructions include landmarks only at decision points, there are fewer intersections drawn on sketch maps. More so, the drawn sketched maps provide a spatial configuration that is hardly recognizable. Therefore, we suggest that providing wayfinding instructions with global landmarks and local landmarks (at decision points and along the route) contribute to cognitive mapping efficiently that a person can acquire reasonable spatial configuration.

6 Conclusion

Besides generating instructions that are easy to follow, our major research interest is addressing cognitively efficient wayfinding instructions that can also facilitate spatial orientation and cognitive mapping. In this study, we investigate the roles of different types of verbal wayfinding instructions on spatial orientation and cognitive mapping.

The most important finding is that including global and local landmarks in route instructions contributes to spatial orientation and cognitive mapping. Landmarks located in distance, at potential decision points, and along a route help a person to acquire reasonable spatial configuration of an environment. This acquired spatial configuration consequently helps a person to better orient in an environment. Despite its supportive role on spatial orientation, this type of instructions does not lead to accurate acquisition of distance information.

Unlike what we previously assumed, the machine-generated instructions, which include distance information for each segment, still remain challenging for a person to acquire spatial knowledge about distance among features in an environment.

Due to the preliminary status of our study, we have not addressed the effects of different types of route instructions on actual wayfinding performance. The results here have provided us promising information that efficiency of wayfinding, spatial orientation and cognitive mapping can be achieved through including global and local landmarks at various locations in route instructions. We are conducting this study with a larger number of participants, which would lead us to a more comprehensive understanding. This study also raises questions including the investigation of the effects of route instructions given in different formats such as map, as well as the generation of orientation-based instructions in an efficient and automatic way. These are the logical follow-ups for us to address in future studies.

Acknowledgements

Research for this paper is based upon work supported by the German Research Foundation (DFG) under grant number SCHW 1372/15-1. The views, opinions, and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the funding agency or the German government.

7 References

[1] Downs, R.M., Stea, D.: Cognitive maps and spatial behavior: Process and products. In: Downs, R.M., Stea, D. (eds.) *Image and Environment*. Aldine, Chicago (1973)

[2] Siegel, A.W., White, S.H.: The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior* 10, 9-55 (1975)

[3] Raubal, M., Winter, S.: Enriching wayfinding instructions with local landmarks. In: Egenhofer, M.J., Mark, D.M. (eds.) *Geographic Information Science: Proceedings of the Second International Conference, GIScience 2002 Boulder, CO, USA, September 25-28,2002*, vol. LNCS2478, pp. 243-259. Springer, Berlin (2002)

[4] Michon, P.E., Denis, M.: When and why are visual landmarks used in giving directions? *Spatial Information Theory, Proceedings of International Conference on Spatial Information Theory*, pp. 292-305 Springer, Berlin (2001)

[5] Tom, A., Denis, M.: Language and spatial cognition: Comparing the roles of landmarks and street names in route instructions. *Applied cognitive psychology* 18, 1213-1230 (2004)

[6] Lovelace, K.L., Hegarty, M., Montello, D.R.: Elements of good route directions in familiar and unfamiliar environments. *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, pp. 65-82 Springer, Berlin (1999)

[7] Schwering, A., Li, R., Anacta, V.J.A.: Orientation information in different forms of route instructions. The 16th AGILE Conference on Geographic Information Science, Leuven, Belgium (2013)

[8] May, A., Ross, T., Bayer, S.H., Tarkiainen, M.J.: Pedestrian navigation aids: information requirements and design implications. *Personal and Ubiquitous Computing* 7, 331-338 (2003)

[9] Allen, G.L.: Principles and practices for communicating route knowledge. *Applied cognitive psychology* 14, 333-359 (2000)

[10] Ross, T., May, A., Thompson, S.: The use of landmarks in pedestrian navigation instructions and the effects of context. *Mobile Human-Computer Interaction-MobileHCI 2004*, pp. 300-304. Springer, Berlin (2004)
[11] Golledge, R.G., Stimson, R.J.: *Spatial behavior: a geographic perspective*. The Guilford Press, New York, London (1997)
[12] Montello, D.R.: Navigation. In: Shah, P., Miyake, A. (eds.) *Cambridge handbook of visuospatial thinking*, pp. 257-294. Cambridge University Press, Cambridge, England (2005)
[13] Hart, R.A., Moore, G.T.: The development of spatial cognition: A review. In: Downs, R., Stea, D. (eds.) *Image and Environment: Cognitive Mapping and Spatial Behavior*, pp. 124-288. Aldine Publishing, Chicago (1973)
[14] Loomis, J.M., Blascovich, J.J., Beall, A.C.: Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods* 31, 557-564 (1999)
[15] Michon, R.-E., Denis, M.: When and why are visual landmarks used in giving directions? In: Montello, D.R. (ed.) *Spatial Information Theory, Proceedings of International Conference on Spatial Information Theory*, pp. 292-305. Springer, Berlin (2001)
[16] Denis, M.: The description of routes: A cognitive approach to the production of spatial discourse. *Cahiers de psychologie cognitive* 16, 409-458 (1997)
[17] Richter, K.-F., Klippel, A.: A model for context-specific route directions. *Spatial Cognition IV. Reasoning, Action, Interaction*, pp. 58-78. Springer, Berlin (2005)
[18] Couclelis, H.: Verbal directions for way-finding: space, cognition, and language. *The construction of cognitive maps*, pp. 133-153. Springer, Berlin (1996)
[19] Steck, S.D., Mallot, H.A.: The role of global and local landmarks in virtual environment navigation. *Presence* 9, 69-83 (2000)
[20] Winter, S., Tomko, M., Elias, B., Sester, M.: Landmark hierarchies in context. *Environment and Planning B: Planning and Design* 35, 381-398 (2008)
[21] Tom, A., Denis, M.: Referring to landmark or street information in route directions: What difference does it make? In: Kuhn, W., et al. (eds.) *Spatial Information Theory. Foundations of Geographic Information Science*, pp. 362-374. Springer Berlin (2003)
[22] Hegarty, M., Richardson, A.E., Montello, D.R., Lovelace, K.L., Subbiah, I.: Development of a self-report measure of environmental spatial ability. *Intelligence* 30, 425-447 (2002)
[23] Lawton, C.: Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles* 30, 765-779 (1994)
[24] Guay, R.: *Purdue spatial visualization test*. Purdue University (1976)

Session:
Mapping and the Citizen Sensor
COST TD1202

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Semantic analysis of Citizen Sensing, Crowdsourcing and VGI

Alexis Comber
University of Leicester
Leicester, UK
ajc36@le.ac.uk

Sven Schade
JRC
Ispra, Italy
sven.schade@jrc.ec.europa.eu

Linda See
IIASA
Laxenburg, Austria
see@iiasa.ac.at

Peter Mooney
NIUM
Maynooth, Eire
peter.mooney@nuim.ie

Giles Foody
University of Nottingham
Nottingham, UK
giles.foody@nottingham.ac.uk

Abstract

This paper describes a semantic analysis of terms used to describe citizen sensing and crowdsourced data use in scientific analyses. It applies a latency analysis to journal abstracts downloaded from Scopus that matched one of number of terms related to crowd sourced data and citizen science. The latency analysis shows how the terms associated with crowdsourcing are related and how they have evolved over time.

1 Introduction

Whilst there is a long tradition of members of the public recording and sharing information about the world we live in, recent developments in digital technologies have driven an explosion of crowdsourced data collection and creation. Due to connected, location enabled digital devices – smartphones, cameras, tablets, notebooks etc – citizens are able to capture and almost share spatially referenced information about all kinds of processes (see for example [3] [6] [8]) via many different types of platforms – the web, social networks, server host sites (e.g. Flickr for photographs) – as well as targeted activities such as OpenStreetMap [9] and Geograph. Thus, it is now relatively simple for citizens to capture and share information about the world they live in, both actively (e.g. via OSM creation) or passively (e.g. via mining of twitter feeds).

The recent high level of scientific interest in crowdsourced data is high for 2 simple reasons. First, the very high data volumes that are potentially available to the scientist, and second, the low cost of such data. That is, at the core of much of the current scientific interest is the possibility that crowdsourced data may be able to replace data collected under the designed experiment that is where data are collected under a formal experimental design that includes sampling strategies, stratifications, etc. However, the critical issue using crowdsourced data in this way relates to the quality of the data. This not only relates to the reliability of observations and their labeling - whether they truly describe the phenomenon under consideration, but also to the spatial distribution of the observations, which depends on the locations of the individuals volunteering the information. Thus the controls over what is recorded and where is recorded that are frequently addressed by pre-specified experimental designs and the establishment of data capture protocols are lacking in crowdsourced data.

The focus of this paper is to consider how conceptualisations of crowdsourced data have evolved over time. It analyses the semantics of 'citizen science' activities as reported in the scientific literature for the period 1990 to 2013 in order to understand the changes in the way that the scientific community use, conceive apply such data.

2 Analysis

A text mining analysis of the semantics used in research describing the analysis, acquisition and qualities of crowdsourced geographic information was undertaken. The abstracts of 10,441scientific papers, published between 1990 and 2013, that contained any of the 24 the terms listed in Table 1 in their title, keywords or abstract were downloaded from Scopus (note: these terms were selected as initial set to investigate – future work will extend and refine these).

Table 1. Search terms used to extract scientific papers form Scopus

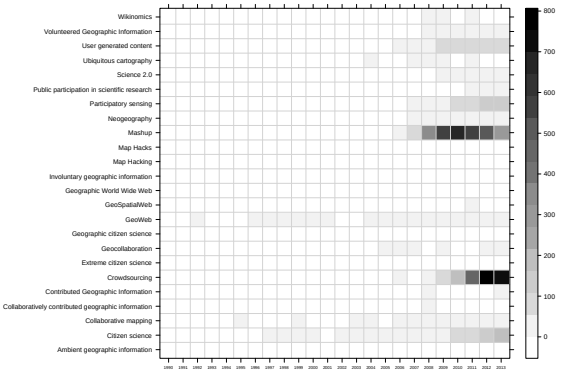
Terms
Science 2.0
Collaborative mapping
Wikinomics
Extreme citizen science
Geographic citizen science
Geocollaboration
Map Hacking or Map Hacks
Neogeography
Participatory sensing
Ubiquitous cartography
Mashup
Citizen science
Collaboratively contributed geographic information
Crowdsourcing
Geographic World Wide Web
GeoWeb or GeoSpatialWeb
Involuntary geographic information
Volunteered Geographic Information
Public participation in scientific research
Ambient geographic information
User-generated content
Contributed Geographic Information

A *Latent Dirichlet Allocation* (LDA), first proposed by [1] was used to analyse the content of the abstracts. LDA seeks to explain similarity in documents using unobserved, latent groups or *topics*. The idea is that each document includes a number of embedded topics which are indicated by the words that the documents contain and that the frequency of words in documents describe these associations. Latent approaches consider the data (documents) and the hidden concepts they contain (topics) from the standpoint of naivety and seek to determine the underlying similarities between documents and concepts. These techniques have been used in a number of spatial data analyses [11] [12] [4] [5] have applied them to integrate land cover data with different taxonomies. Here, citation data were downloaded from Scopus for publications that matched at least one of a number of search criteria.

The data were cleaned to remove English stopwords (conjunctions, pronouns etc.), numbers, punctuation, whitespaces and any words less than 3 characters long. The words were then *stemmed*. Stemming is the process of establishing common etymological roots for words such that, for example *propose* and *proposal* have the same stem of *propos*. The cleaned and stemmed abstracts were then organised into a *corpus* of 24 documents based on the year of publication.

The evolution of the terms and phrases related to citizen sensing listed above was analysed using the *term frequency-inverse document frequency* (tf.idf). The tf.idf weight is a commonly used in library sciences for document classification and information retrieval. It is a statistical measure and describes the importance of a word in relation to any given document. A frequency matrix was constructed describing the occurrence of each of the phrases in each of the 24 documents representing the corpus of abstracts for each year (1990 to 2013). This is shown in Figure 1 where the cells in the matrix indicate the number of times each term appears in each year. Note, that in this case corpuses were re-created for each year, no stemming or removal of stop words was performed, and search terms with more than one word were replaced with concatenated versions (e.g. such that "Citizen science" was replaced with "Citizen_science").

Figure 1: The frequency of occurrence for each search term.

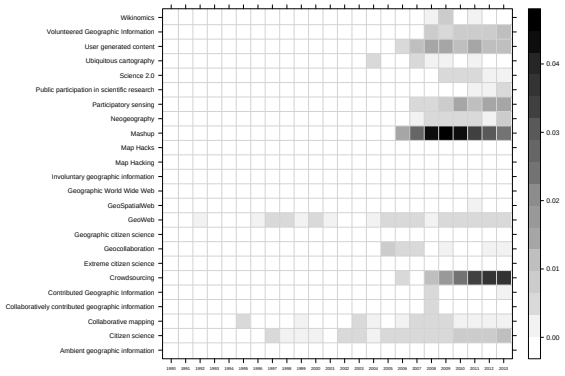


The terms in the matrix were weighted using the 'tf.idf' scheme described in [9]:

$$W_{ij} = \frac{n_i}{\sum n_i} ln \frac{D}{n_i}$$

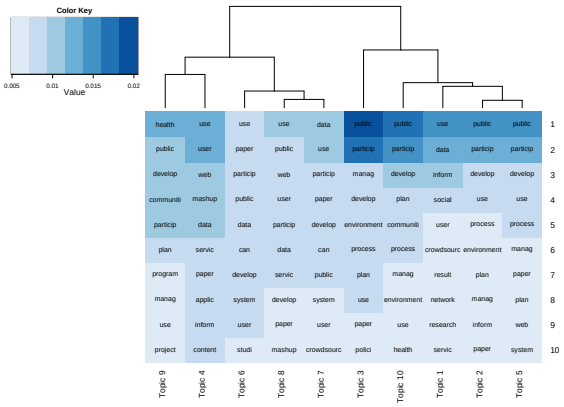
where W_{ij} is the weight of the i^{th} word in the j^{th} class, n_i is the number of times the word appears in the j^{th} class, Σn_i is the total length of the j^{th} class description, D is the total number of classes and n_j is the number of classes containing the i^{th} word. The weighting has the effect that a word that appears in all class descriptions has a zero weight, but a word appearing frequently in a few short classes has a high weight. The results of apply the are shown in Figure 2.

Figure 2. The changes in *tf.idf* values for the search terms 1990 to 2013.



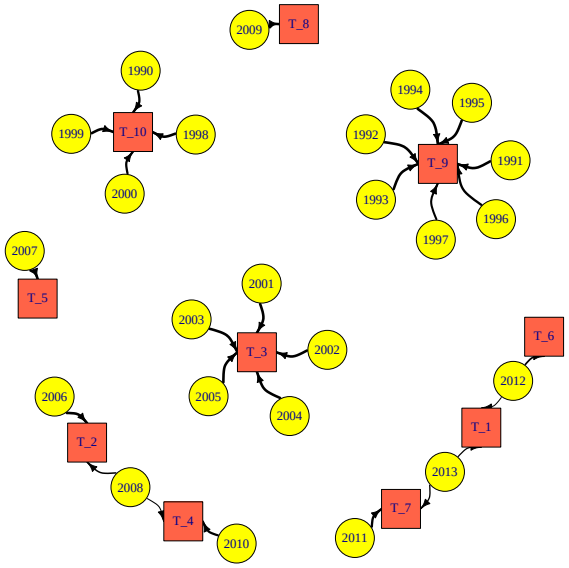
A Latent Dirichlet Allocation analysis was run on the corpus using the *topicmodel* package [2] in R, the opensource statistical software. Ten latent variables or topics were identified and these can be characterised by the terms that are most strongly associated with them from the posterior probabilities generated by the LDA of each term being associated with each topic (Figure 3). This suggests that there are 3 distinct topic groups: Topics 4 and 9 (*community*, *mashup*, *web*, *develop*, *health*) , Topics 6, 7 and 8 (*use*, *particip*, *web*, *public*) and Topics 1, 2, 3, 5 and 10 (*particip*, *develop*, *public*).

Figure 3. The 10 stemmed terms most strongly associated with each topic, shaded by the posterior probability of belonging to that topic and with the topics clustered.



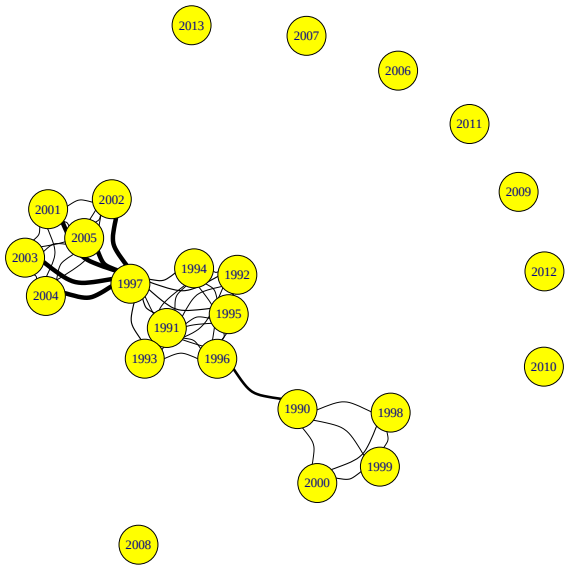
The LDA also generates posterior probabilities that each document is associated with each topic. These relationships between topics and documents via their semantics for each year can be visualised in a network, where the edges are defined by probability. For clarity the edges (connections) between years and topics (vertices) were removed if the posterior probability for each topic-year pair was less than the minimum posterior probability for that year plus the standard deviation [7]. The connections between topics and years is shown in Figure 4 and indicates an evolution over time of the concepts associated with publications in this domain.

Figure 4. The links between topics and years, with the strength of the link as defined by the posterior probability as determined by the LDA model indicated by the edge widths.



The connectedness between the semantics embedded in documents from different years is further illustrated in Figure 5. This shows the semantic distances between the documents for different years in the corpus. The recent explosion of publications, application and the wider discussion of the use of citizen sensed data in scientific publications are perhaps suggested by the lack of links between publications from more recent years compared to the 1990s and early 2000s – 1997 is particularly interesting year.

Figure 5. A network describing the semantic distances between documents published in different years.



3 Discussion Points

A number of areas for future consideration have been identified through this initial exploratory work. First, that the number of scientific papers that cite (not about) crowdsourcing topics has increased in recent years. Second that there are clearly identifiable evolutionary phases in the way that such information is referred to, witness the links in Figure 4 and Figure 5. These potentially reflects phases in GIS Science related to crowdsourcing between 1990 and 2005, the beginning of mashups, neogeography and so on in 2005-2006 seeing and a breadth of citizen science activities since then appearing to be disconnected. Thirdly, that recent research is clearly drawing from a much wider range of data sources, labelled in different and novel ways, potentially reflecting the rapid increase in the platforms and systems available to individual citizens that enable them capture and share a diverse range of different types of information, describing the world we live in. There are obvious areas for future research in considering who contributes such data, the impact of digital divides on the nature of the information that is contributed and potential biases towards western, developed populations and of course the nature of the technologies used to capture and share such information. On-going work is considering these issues

4 Acknoweldgements

This work was undertaken under the EU COST TD1202 ‘Mapping and the citizen sensor’.

References

[1] Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

[2] Grun, B and Hornik, K, 2013. Package 'topicmodels'. http://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf [available 07/01/14]

[3] Coleman, D., 2010. The potential and early limitations of volunteered geographic information. *Geomatica* 64 (2), 27–39.

[4] Comber, A.J., Fisher, P.F. and Wadsworth, R.A., (2007). Mining semantics of geographical information to generate user-relevant metadata, Spatial Data Usability Workshop, at *AGILE 2007, 10th AGILE conference on Geographic Information Science*, (eds. Monica Wachowicz and Lars Bodum), 8th May, Aalborg.

[5] Comber, A, Lear, A and Wadsworth, R, (2010). A comparison of text mining and semantic approaches for integrating national and local habitats data: semantic accuracy, error or inconstancy? In proceedings of Accuracy 2010. Pp297-300 in N Tate and P Fisher (eds.), *Proceedings of the 9th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 20th -23rd July 2010, University of Leicester, Leicester.

[6] Haklay, M., Basiouka, S., Antoniou, V., Ather, A., 2010. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. Cartographic Journal 47, 315–322.

[7] Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

[8] Jones, C.E., Mount, N.J., Weber, P., 2012. The rise of the GIS volunteer. *Transactions in GIS* 16, 431–434.

[9] Mooney, P., Corcoran, P., 2012. The annotation process in OpenStreetMap. *Transactions in GIS* 16, 561–579

[10] Robertson, S.E. and Spärck Jones, K., 1976, Relevance weighting of search terms, *Journal of the American Society for Information Science*, 27(3), 129-46.

[11] Wadsworth, R.A, Comber, A.J. and Fisher, P.F., (2008). Probabilistic Latent Semantic Analysis as a potential method for integrating spatial data concepts, pp 99-108 in *Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday*, (ed. Gerhard Navratil), GeoInfo Series 39, Vienna, ISBN 978-3-901716-41-6

[12] Wadsworth RA, Comber AJ, and Fisher PF, 2006 Expert knowledge and embedded knowledge: or why long rambling class descriptions are useful. In *Progress in Sptial Data Handling, Proceedings of SDH 2006*, (eds Andreas Riedl, Wolfgang Kainz, Gregory Elmes), Springer Berlin: 197 – 213

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Characteristics of Citizen-contributed Geographic Information

Spyridon Spyratos
University of Thessaly,
Pedion Areos, 38334, Volos, Greece
&
Joint Research Centre, European Commission,
via Enrico Fermi 2749, 21027, Ispra, Italy
spyridon.spyratos@jrc.ec.europa.eu

Michael Lutz
Joint Research Centre,
European Commission,
via Enrico Fermi 2749,
21027, Ispra, Italy
michael.lutz@
jrc.ec.europa.eu

Francesco Pantisano
Joint Research Centre,
European Commission,
via Enrico Fermi 2749,
21027, Ispra, Italy
francesco.pantisano@
jrc.ec.europa.eu

Abstract

Current Internet applications have been increasingly incorporating citizen-contributed geographic information (CCGI) with much heterogeneous characteristics. Nevertheless, despite their differences, several terms are often being used interchangeably to define CCGI types, in the existing literature. As a result, the notion of CCGI has to be carefully specified, in order to avoid vagueness, and to facilitate the choice of a suitable CCGI dataset to be used for a given application. To address the terminological ambiguity in the description of CCGI types, we propose a typology of GI and a theoretical framework for the evaluation of GI in terms of data quality, number and type of contributors and cost of data collection per observation. We distinguish between CCGI explicitly collected for scientific or socially-oriented purposes. We review 27 of the main Internet-based CCGI platforms and we analyse their characteristics in terms of purpose of the data collection, use of quality assurance and quality control (QA/QC) mechanisms, thematic category, and geographic extents of the collected data. Based on the proposed typology and the analysis of the platforms, we conclude that CCGI differs in terms of data quality, number of contributors, data collection cost and the application of QA/QC mechanisms, depending on the purpose of the data collection.

Keywords: Volunteered Geographic Information (VGI), Citizen Science, Crowd sourced geographic information, Citizen-Contributed Geographic Information (CCGI), Social Geographic Data (SGD)

1 Introduction

Recent social and technological developments, such as the increased educational attainment and the diffusion of sensor-enabled devices increase the number of citizens who are potentially able to collect and publicly share almost real time geographic information (GI) on the Internet. Such a citizen-contributed geographic information (CCGI) differs from GI collected by professionals in the context of professional routines and practices for four main reasons. First, the CCGI data collectors possess significantly diverse level of scientific and technical knowledge [2]. Second, the CCGI data collection methods and equipment are very different and often unknown. Third, the quality of CCGI is not always ensured and controlled by formal quality assurance procedures [14], and, finally, CCGI is mostly collected at time and locations that are generally not defined a priori by an organization.

Lately, an increasing number of Internet-based platforms has been developed with the purpose of collecting CCGI for both socially-oriented and scientific purposes. These platforms consist of hardware and software components, such as servers and mobile application interfaces, as well as analytical tools for data processing. They cover data about various environmental domains, such as acoustic pollution [30], biodiversity [16] and land cover observations [8]. Clearly, since CCGI data is gratuitously contributed by the citizens, these platforms offer timely GI and at very limited cost [11].

Due to these reasons, CCGI is increasingly used as auxiliary input for environmental monitoring and mapping [20, 29] and research studies [7]. However, due to the numerous types of existing CCGI, it is still unclear whether and what types of CCGI can contribute towards a better and more holistic understanding of the environment. Goodchild and Li [11] suggest that volunteered geographic information (VGI) is often inadequate data source for scientific research, because "its quality is highly variable and undocumented, it fails to follow scientific principles of sampling design, and its coverage is incomplete". In contrast, Lee [18] mentions that much of the knowledge about the USA climate is based on long-term volunteer records. In this respect, we argue that both of the above statements are valid, as they refer to different types of CCGI.

In fact, CCGI is not a homogenous category and includes GI that significantly differs in terms of purpose of data collection, data quality and the characteristics of contributors. Nevertheless, in the literature, terms such as VGI [10], crowd sourced geographic information, and user generated geographic content (UGGC) are often being used interchangeably to describe various GI types. For example, VGI describes a distinct subset of CCGI, UGGC and crowd-sourced GI as it embodies the notion of volunteering for data collection [5]. VGI describes a science-oriented phenomenon

that is supported by technology. Devising CCGI categories is a fundamental operation, as the definition of each of these categories has to denote the characteristics of the collected data, and the characteristics of the contributors e.g. volunteers or users of social networking applications.

In this study, we address this terminological ambiguity in the description of CCGI types, and we provide guidelines for GI type definition. First, based on the purpose of the data collection activity, we propose a typology of CCGI and we identify factors that affect the data quality and quantity of the collected data. Second, we identify Internet-based platforms that collect CCGI, we classify them based on the proposed typology, and we analyse three characteristics of CCGI platforms and datasets. These characteristics are: (a) the existence of quality control and quality assurance (QA/QC) mechanisms that depend on citizens, (b) the thematic category, and (c) the geographic extent of the collected data.

The main rationale of this work is to propose a theoretical framework for the evaluation of CCGI data to be used for scientific or social applications.

The remainder of the paper is structured as follows. Section 2 describes the proposed typology of CCGI. Section 3 presents the methodology followed for identifying and analysing CCGI platforms and datasets and the results of their analysis. In Section 4, we discuss the results of the analysis. Finally, future work and conclusions are outlined in Section 5.

2 Typology of citizen-contributed geographic information

The existing literature includes two CCGI typologies [1, 3] which relevant to the purpose of the current study. The first, proposed by Antoniou et al. [1], introduces a distinction between spatially implicit and explicit UGGC web applications, based on their declared objectives. The second, by Craglia et al. [3], defines four VGI types based on two dimensions which can be either explicit or implicit. These dimensions are "first, the way the information was made available, and second, the way geographic information forms part of it" [3].

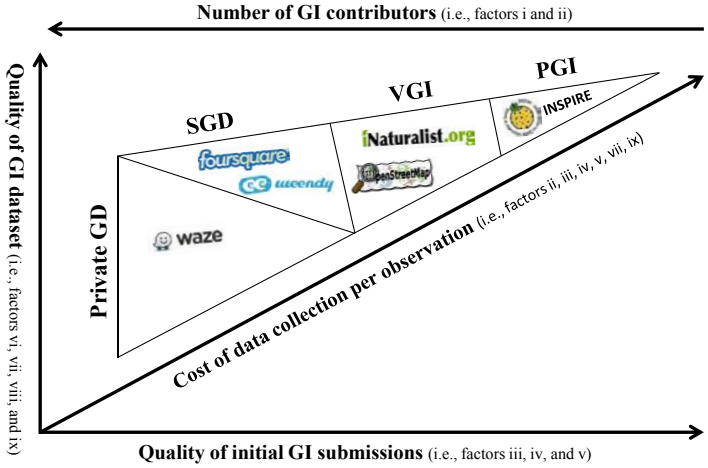
To address the terminological ambiguity in the description of CCGI types, and to support the analysis of platforms, provided in Section 3, we propose a typology of GI which, in contrast to the existing ones, is based on the purpose of the data collection. In the proposed typology (see Fig. 1) we distinguish between CCGI collected for scientific (VGI) and socially-oriented (Social Geographic Data) purposes which are defined as:

- Volunteered Geographic Information (VGI). In this study VGI refers to GI intentionally collected by citizens, in the context of real life or on-line science-oriented voluntary activities. For instance, the VGI category includes GI collected by volunteers as part of a broad scientific enquiry in the data collection stage of citizen science projects (for more details on citizen science see Silvertown [25]) or in the context of crowdsourcing projects [15] e.g. Google Map Maker [12].
- Social Geographic Data (SGD). The SGD category describes geographic or geo-referenced data that is publicly available over the Internet and it has been generated by citizens for socially oriented purposes. For example, this category includes Foursquare place data [6], and geo-located public tweets [28].

Apart from the above CCGI types, two other categories of GI exist:

- Professional Geographic Information (PGI) [22]. PGI is composed by GI exclusively collected by experts, e.g. surveyors or urban planners, in the context of professional routines and practices.
- Private Geographic Data (Private GD) category includes geographic or geo-tagged data that has not been publicly shared by the data author. Private GD is produced by citizens and it can either be data that is associated with the characteristics of an individual or data intended for a particular person, group or service. For example, this category includes not-publicly shared geo-located tweets [28], and Global Navigation Satellite System (GNSS) data contributed to navigation services.

Fig. 1: Typology of GI



This paper focusses on CCGI, i.e., GI collected and publicly shared by citizens. PGI and Private GD are out of the scope of this study, since the former includes only qualified professional in its collection, and the latter deals with data not publicly contributed and not intended to be reused, other than by the initial recipients.

2.1 Characteristics of GI datasets

In the proposed typology, we distinguish between three main characteristics (see Fig. 1) for SGD, VGI, PGI and Private GD. The characteristics of the data collection activity, of the GI contributors, platforms and data collection tools, are factors that impact the characteristics of the collected data. These characteristics are: *the number of potential GI contributors*, *the quality of initial GI submission*, *the overall quality of the GI datasets*, and *the cost of data collection*. Due to the scope of this study the analysis is focused on the CCGI, namely the SGD and the VGI.

2.1.1. Number of potential GI contributors

As shown in the upper axes of Fig. 1, the number and the demographic profile of citizens that can potentially collect GI depends on the following factors:

i. The level of technical and scientific knowledge required for data collection.
ii. The time, technical equipment and other resources needed for data collection [13].

These two factors limit the number of citizens who can autonomously participate in science-oriented or socially-oriented data collection activities. Regarding the scientific and technical knowledge of VGI data collectors (i.e. factor i), a study by Budhathoki et al. [2] revealed that 25% of the OpenStreetMap contributors had more than 1 year experience with GISystems and the 49% had none. Statistics like these highlight the fact that the demographic profile of VGI data collectors is heterogeneous and not representative of the society. Additionally, such statistics prove that VGI data collectors are not largely untrained, and confirm Lee's [18] statement that volunteer does not necessarily equal amateur.

In contrast to VGI, SGD is not the product of science-oriented tasks, and thus, the level of scientific knowledge required for the collection of SGD observations is, in principle, lower compared to VGI. Thus, SGD can additionally be collected by citizens with low-level science skills. As a result, the number of potential SGD contributors is typically larger compared to the number of VGI contributors.

2.1.2. Quality of initial GI submissions

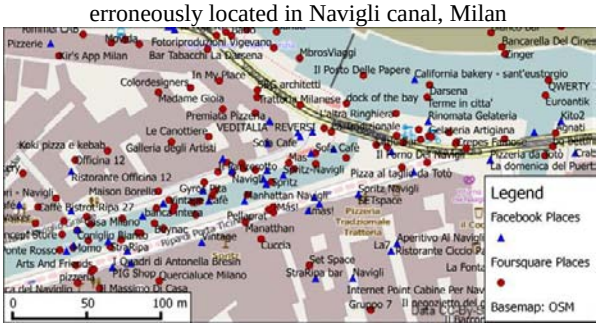
The quality of initial GI submissions refers to the quality of the first GI data submission by a citizen, before any correction or filtering is made by the QA/QC mechanisms. For an extensive survey on the quality elements of GI, such as the positional and thematic accuracy, we refer the interested reader to Oort [21]. As shown in the bottom axes of Fig. 1, the quality of initial GI submissions depends on factors such as:

iii. The desired (or de-facto, de jure) accuracy of GI.
iv. The scientific and technical knowledge of data collectors [4, 24].
v. The accuracy of the utilized equipment, sensors, and auxiliary data, e.g. satellite images.

Factor (iv) relies on the contributors characteristics, while factors (iii), and (v) also depend on the platforms. For instance, for mapping applications, the accuracy of an observation depends both on the accuracy of the GNSS sensors that citizens deploy, and on the quality of the auxiliary satellite images that a platforms provides.

According to our definition, VGI is collected for scientific purposes, and thus, the desired positional and thematic accuracy (i.e. factor iii) and the quality of utilized sensor (i.e. factor v) are both higher compared to SGD. The reason is that a volunteer aims at describing a phenomenon or a feature as accurately as possible. Instead, users of socially-oriented web applications demand a level of accuracy that is sufficient to efficiently convey a geo-tagged message. For example, Fig. 2 shows the Navigli area in Milano, Italy, where many of the Facebook and Foursquare places are mistakenly pinned in the water. The place data positional precision is clearly not suitable for mapping or routing purposes.

Fig. 2: Many Facebook and Foursquare place data are erroneously located in Navigli canal, Milan



Sources: Place data, Facebook Graph API and Foursquare Venues API; Basemap, OSM contributors.

2.1.3. Quality of GI datasets

The quality of VGI and SGD significantly varies across time and space, even within the same dataset. As a matter of fact, VGI and SGD datasets are highly heterogeneous, as they are composed by observations that differ in terms of equipment accuracy and citizen technical and scientific background, even in local spatial scale. We note that the overall quality of the GI datasets in a given area mainly depends on the following factors:

vi. The quality of the initial GI submissions.
vii. The number and the demographic profile of contributors and the number of contributions.
viii. The existence and the application of QA/QC mechanisms.
ix. The degree of coordination for the data collection activity.

The quality of GI datasets is determined to a great extent by the quality of initial GI submissions (i.e. factor vi) from which are derived. The demographic profile, the number and the

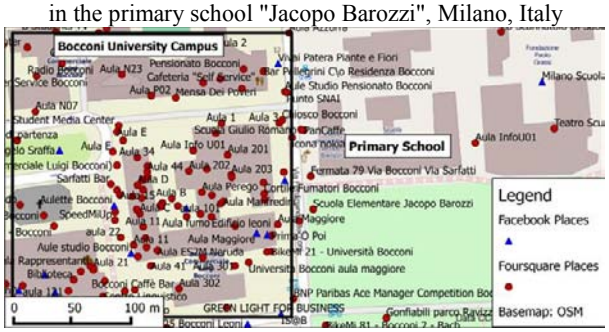
spatial distribution of CCGI contributors are factors (factor vii and in more detail see Section 2.1.1) that affect the thematic and spatial completeness of a CCGI dataset [13, 27]. The existence of horizontal or hierarchical coordination of a data collection activity (i.e. factor ix) clearly has a positive impact on the spatial and temporal completeness of a dataset.

QA/QC mechanisms are adopted for the purpose of improving the quality of GI. QA/QC mechanisms can be managed by professionals in the context of professional routines and practices, and/or by the community of contributors, in case citizens assess the correctness of the observations. In addition, QA/QC mechanisms can be supported by automated procedures, in which each observation is automatically checked based on predefined rules, as in [19], for example. In citizen-based QA/QC mechanisms, the quality of the observations stored in the GI datasets depends on the number of contributors (i.e. factor vii), which are also reviewers [9, 14]. This relation directly confirms "Linus' law" [23], stating that the higher the number of users or contributors of a product is, the higher is the probability that a problem will be fixed by someone.

Several studies have proved that the overall quality of VGI datasets is inferior to PGI [9, 13, 17]. However, few studies have addressed the quality of SGD. The Antoniou e. al. [1] study demonstrate that the spatial distribution of SGD observations is more likely to be limited to the users' existing activity space compared to VGI spatial distribution. SGD is collected in the context of the data collectors' social activities, and not as part of a scientific inquiry. For this reason, VGI datasets are expected to have higher spatial and temporal completeness, compared to SGD.

For instance, Fig. 3 shows Foursquare and Facebook place data in an area of Milan, Italy. On the left side of Fig.3, the Bocconi University is well covered while a primary school on the right side is not. The reason for this is that only a limited number of primary school students or staff are declaring the physical presence on Facebook or Foursquare. As a result, their activity space is not well covered on Facebook and Foursquare place datasets.

Fig. 3: Abundance of Facebook and Foursquare place data in a detailed level in Bocconi University campus at the left side of the figure, versus scarcity of place data at the right side, e.g., in the primary school "Jacopo Barozzi", Milano, Italy



Sources: Place data, Facebook Graph API and Foursquare Venues API; Basemap, OSM contributors.

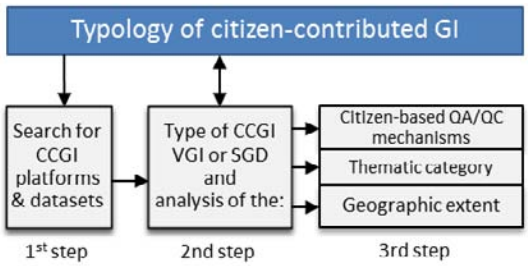
2.1.4. Cost of data collection per observation

The financial cost of data collection and processing per observation is another important characteristic of GI. Factors that affect this financial cost are ii, iii, iv, v, viii, and ix. In principle, the higher the quality of the technical and human resources used for data collection, the higher the cost for their usage is. For example, professional GNSS receivers are more accurate and expensive than those built-in mobile phones [31]. The application of QA/QC mechanisms, and the efforts made for coordination of the data collection activity are also factors that have a considerable financial cost for data collection. As a matter of fact, each GI type incurs different costs for data collection. For the collection of PGI, a professional staff is hired, while for the VGI and SGD the contributors are volunteers. Professional trainers are commonly used to train PGI and VGI data collectors, while this is not the case for SGD and Private GD. It is, therefore, arguable that SGD is less expensive to collect than VGI and PGI.

3 Methodology & Results

In this section, we focus our analysis on Internet-based platforms that collect CCGI about environmental elements, such as atmosphere, water, soil, land and landscape. We decided to analyse CCGI platforms, in an effort to study how the purpose of the data collection affects the characteristics of the collected CCGI datasets. The methodology for identifying and analysing CCGI platforms and datasets is presented in Fig. 4.

Fig. 4: Methodology followed for identifying and analysing CCGI platforms and datasets



The first step of the methodology was the identification of CCGI platforms that collect data on the environmental elements. For the identification of these platforms an extensive search of the English literature and Web resources was conducted. The searches were performed by using English keywords, which are typically used to describe CCGI. These terms and their variants are:

a) Volunteered geographic/environmental information/data
b) User-generated geographic/spatial content.
c) Crowd sourced geographic/environmental information/ data.

During the search period, 27 platforms (see Table 1) were identified. Given the method for identifying the platforms, the results mostly include popular English-based platforms. Therefore, the results of the platforms analysis cannot be quantitatively generalized, but could be used for understanding the CCGI characteristics.

The second step of the methodology was the analysis of the type of CCGI that the 27 platforms collect. Based on the proposed typology, we classified the 27 platforms into VGI and SGD (see third column of Table 1). The reason for this is that the purpose of the data collection, as defined by each platform's objectives, affects the characteristics of the collected data, such as, its spatial distribution and its accuracy.

Table 1: Name and type and website of CCGI platforms

No	Name of platform	Type of GI	Website
1	Aircasting	VGI	aircasting.org
2	AirProbe	VGI	cs.everyaware.eu/event/airprobe
3	ARGO Sentinel	VGI	argomobile.isti.cnr.it
4	CWOP	VGI	wxqa.com
5	Facebook Places	SGD	www.facebook.com
6	FishBase	VGI	www.fishbase.org
7	Flickr	SGD	www.flickr.com
8	Foursquare Venues	SGD	foursquare.com
9	Geograph	VGI	www.geograph.org.uk
10	Geowiki	VGI	www.geo-wiki.org
11	Google Map Maker	VGI	www.google.com/mapmaker
12	iNaturalist	VGI	www.inaturalist.org
13	iRecord	VGI	www.brc.ac.uk/irecord
14	iSPEX	VGI	ispex.nl/en
15	iSpot	VGI	www.ispot.org.uk
16	NoiseTube	VGI	www.noisetube.net
17	Noisewatch	VGI	eyeonearth.org/map/NoiseWatch
18	OpenStreetMap	VGI	openstreetmap.org
19	Panoramio	SGD	www.panoramio.com
20	PSW Weather	VGI	www.pwsweather.com
21	The National Map Corps	VGI	navigator.er.usgs.gov
22	WaterWatch	VGI	eyeonearth.org/map/waterwatch
23	Weathersignal	VGI	weathersignal.com
24	WeatherUnderground	VGI	www.wunderground.com
25	Weendy	SGD	www.weendy.com
26	Wheel Map	VGI	www.wheelmap.org
27	WideNoise	VGI	cs.everyaware.eu/event/widenoise

The third step of the methodology included the analysis of three characteristics of CCGI platforms and datasets. The first characteristic that we analysed is the type of QA/QC mechanisms which depend on citizens. Citizen-based QA/QC mechanisms allow the users of the platforms to review and rate the correctness of VGI and SGD observations. Citizen-based QA/QC mechanisms can vary from being horizontally structured, in which user have distributed and equal authorities on editing observations, to more hierarchically structured, in which community representatives or elite users have increased editing authorities compared to average users.

There are two types of citizen-based QA/QC mechanisms. The first type allows citizens to edit an observation or suggest an edit to its author. The second type allow citizens to rate the accuracy of an observation, and thus, to also assess the competence of the data contributor. Based on the existence and the type of citizen-based QA/QC mechanisms, we classified the 27 platforms in four categories (see Table 2).

Table 2: Citizen-based QA/QC mechanism of CCGI platforms

Citizen-based QA/QC	VGI Platforms	SGD platforms
None	1; 2; 3; 4; 13; 14; 16; 17; 20; 22; 23; 24; 27	7; 19; 25
Only rate	None	None
Only edit	6; 9; 10; 11; 18; 21; 26	5; 8
Rate and edit	12; 15	None

The second characteristic that we analysed was the thematic category of the data that the 27 platforms collect. We used a context based classification into six thematic categories as shown in Table 3. Moreover, we classified the six thematic categories into two groups. The first includes CCGI about continuous geographic phenomena and the second CCGI about discrete geographic features.

Table 3: Thematic category of CCGI datasets

	Thematic category	VGI platforms	SGD platforms
Phenomena	Noise	1; 16; 17; 27	None
	Meteorology	1; 2; 4; 20; 23; 25	25
	Air quality	1; 2; 14	None
	Water quality	3; 22	None
Features	Biodiversity, species occurrences	6; 12; 13; 15;	7; 19
	Topography, place, land cover and landscape	9; 10; 11; 18; 21; 26	5; 7; 8; 19

Finally, we analysed the geographic extent of CCGI datasets. The geographic extent can be local, national, multi-national, or global. As shown in the Table 4, the geographic extent of the most CCGI data sources that were identified in this study is global.

Table 4: Geographic extent of CCGI datasets

Geographic extent	VGI Platforms	SGD platforms
Global	1; 2; 3; 4; 6; 10; 11; 12; 15; 16; 17; 18; 20; 23; 24; 26; 27	5; 7; 8; 19; 25
Multi-National	9 (UK, IL); 22 (EU)	None
National	13(UK); 21(US); 14(NL)	None
Local	None	None

4 Discussion

SGD and VGI are collected in the context of socially and science oriented activities respectively. As we have discussed in Section 2, SGD and VGI differ in terms of the quality of initial GI submissions, the overall quality of GI datasets, the

number of the potential contributors, and the data collection cost per observation. Although SGD is collected for socially-oriented purposes, it can be reused in the context of scientific applications. An example is given by the reuse of Panoramio photos as auxiliary input for land cover mapping [27].

Most of the VGI platforms and all of the SGD platforms analysed in this study have global geographic extent. The reason is that the identified platforms are biased towards popular, and due to their popularity are more likely to be used by users and volunteers worldwide. The development and maintenance of CCGI platforms is a task that requires significant financial resources and technical skills. Thus, local participatory data collection and citizen science initiatives are more likely to use existing well-established CCGI Internet platforms for collecting data instead of developing new platforms.

As an outcome of the analysis, all the identified CCGI platforms that collect data on continuous geographic phenomena do not consider citizen-based QA/QC mechanisms. Geographic phenomena have properties that change much rapidly. Hence, these observations cannot be easily assessed or edited by other users, as long as they cannot be compared to spatial and temporal near observations of known quality. On the contrary, all the VGI platforms and two SGD platforms, which collect data about geographic features have citizen-based QA/QC mechanisms. The existence of QA/QC mechanisms is enabled by the fact that GI about features can easily be reviewed by citizens that either observe them at a later time, or they re-interpreter a representation of them e.g. images of plants.

Citizen-based rating mechanisms have different purposes in VGI and SGD datasets. The rating of VGI observations is mostly referred to the VGI thematic and positional accuracy, while the rating of SGD observation to their attractiveness/ likability. SGD observations are associated with the subjective perception of citizen about features and phenomena. This provides new research opportunities but it also highlights two important issues. First the statistical representatively of the collected data and second the transparency in the SGD production. The opportunity to include perceptions from contributors could also be evidence of a mixing of quantitative and qualitative information that previous research agendas had called for [26].

5 Conclusions and future work

With the emergence of new Internet applications and mobile devices with numerous embedded sensors, an increasing number of citizens is enabled to potentially contribute various types of GI. Additionally, Internet-based platforms originally meant for socially-oriented purposes are expected to contain more types of geographical, environmental or geo-referenced information, such as weather-tagged photos and messages.

With the plethora of CCGI sources, the selection of a dataset, that fits the data quality requirements (i.e., fitness for use), is a task not always feasible, due to the absence of information on CCGI dataset's quality. Moreover, an on-demand assessment of the CCGI datasets quality is not always possible when reference data of known quality is not available or accessible. The use of CCGI datasets that have not been evaluated in terms of spatio-temporal accuracy and completeness might result in a partial or erroneous understanding of the environment.

In this paper, we have provided a theoretical framework for the evaluation of GI with special emphases on CCGI. Depending on the requirements of an application or research study, and once the proposed framework is fully developed and validated, one will be able to select the type of GI i.e. VGI, SGD, Private GD or PGI, that match the required dataset quality and cost. Moreover, by reviewing the characteristics of the GI collection activity, of the data contributors, platforms and data collection tools, which are listed in Section 2, one can have an indication of the expected accuracy and the spatial distribution of the collected data.

In future work, we will address the validation of the proposed typology. To this end, we will examine VGI and SGD datasets in order to measure the relation between the purpose of data collection and the quality and the cost of the collected data.

Acknowledgement

We would like to thank our colleagues and two anonymous reviewers for their valuable comments and suggestions.

References

[1] V. Antoniou, J. Morley and M. Haklay. Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1):99–110, 2010.

[2] N. Budhathoki, M. Haklay and Z. Nedovic-budic. Who map in OpenStreetMap and why? *Presentation at the State of the Map conference*, Atlanta, USA, 2010.

[3] M. Craglia, F. Ostermann and L. Spinsanti. Digital Earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth*, *5*(5):398–416, 2012.

[4] D. G. Delaney, C.D. Sperling, C.S. Adams and B. Leung. Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biological Invasions*, 10(1):117–128, 2007.

[5] S. Elwood, M.F. Goodchild and D.Z. Sui. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, 102(3):571–590, 2012.

[6] Foursquare. (2014). Venues Platform. Retrieved March 04, 2014, from https://developer.foursquare.com/overview/venues.htm

[7] S. Fritz, L. See, M. van der Velde, R. A. Nalepa, C. Perger, C. Schill, … M. Obersteiner. Downgrading recent estimates of land available for biofuel production. *Environmental Science & Technology*, 47(3):1688–94, 2013.

[8] Geo-Wiki. (2013). The Geo-Wiki Project. Retrieved September 02, 2013, from http://www.geo-wiki.org/

[9] J.-F. Girres and G. Touya. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4):435–459, 2010.

[10] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[11] M. F. Goodchild and L. Li. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110–120, 2012.

[12] Google. (2013). What is Google Map Maker? Retrieved September 05, 2013, from https://support.google.com/mapmaker/answer/157176?hl=en&ref_topic=1093469&rd=1

[13] M. Haklay. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B Planning and Design*, 37(4):682–703, 2010.

[14] M. Haklay, S. Basiouka,V. Antoniou and A. Ather. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. *Cartographic Journal*, 47(4):315–322, 2010.

[15] J. Howe (2006). Crowdsourcing: A Definition. Crowdsourcing. Retrieved July 26, 2013, from http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

[16] INaturalist. (2013). iNaturalist homepage. Retrieved October 24, 2013, from http://www.inaturalist.org/

[17] T. Koukoletsos, M. Haklay and C. Ellul. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, 16(4):477–498, 2012.

[18] V. Lee. Volunteer monitoring: a brief history. *The Volunteer Monitor*, 6(1):29–33, 1994.

[19] NBN. (2013). New NBN Record Cleaner Rules now available. Retrieved November 08, 2013, from http://www.nbn.org.uk/News/Latest-news/New-Record-Cleaner-Rules-now-available.aspx

[20] NoiseWatch. (2013). About NoiseWatch platform. Retrieved July 20, 2013, from http://eyeonearth.org/map/NoiseWatch/

[21] P. V. Oort. Spatial data quality: from description to application. *Wageningen Universiteit*. PhD thesis, 2006.

[22] C. J. Parker, A. May and V. Mitchell. Understanding Design with VGI using an Information Relevance Framework. *Transactions in GIS*, 16(4):545–560, 2012.

[23] E. S. Raymond. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary* (Revised Ed.). O'Reilly Media, Inc. 2001.

[24] L. See, A. Comber, C. Salk, S. Fritz, M. van der Velde, C. Perger, … M. Obersteiner. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS One*, 8(7), 2013.

[25] J. Silvertown. A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9):467–71, 2009.

[26] R.S. Smith. Participatory Approaches Using Geographic Information (PAUGI): Towards a Trans-Atlantic Research Agenda. *In 5th AGILE Conference on Geographic Information Science*, Palma, Spain, 2002.

[27] M. Stephens. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, 78(6):981-996, 2013.

[28] Twitter. (2014). About public and protected Tweets. Retrieved March 04, 2014, from http://support.twitter.com/articles/14016-about-public-and-protected-tweets#

[29] USGS. (2013). Crowd-Sourcing the Nation: Now a National Effort. Retrieved September 09, 2013, from http://www.usgs.gov/newsroom/article.asp?ID=3664#.UihUkD_LI48

[30] WideNoise. (2013). WideNoise Homepage. Retrieved September 03, 2013, from http://cs.everyaware.eu/event/widenoise

[31] P. A. Zandbergen and S.J. Barbeau. Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. *Journal of Navigation*, 64(03):381–399, 2011.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Is this Twitter Event a Disaster?

André Dittrich
Institute of Photogrammetry and Remote Sensing, KIT
Karlsruhe, Germany
andre.dittrich@kit.edu

Christian Lucas
Institute of Photogrammetry and Remote Sensing, KIT
Karlsruhe, Germany
christian.lucas@kit,edu

Abstract

Social media services such as Twitter have become an important channel for reporting real-world events. For example, they can describe the current situation during a disaster. The decisions in crises management are based on detailed on-site information such as what is happening, where and when an event is happening, and who is involved. Thus, in real applications, monitoring the events over social media will enable to analyse the current overall situation. In this paper, the authors introduce a prototype for real-time Twitter-based natural disaster detection and monitoring. The detection approach is multilingual and calculates a statistical based probability for a potential disaster event. For an automatic geo-referencing of the disaster, the approach applies spatial gridding. On this basis the grid cells are subject to a spatial-thematic clustering which uses a method similar to region growing. The application's output is an automatically generated email alert, containing specific information on the disaster.

Keywords: social media monitoring, event detection, real-time analysis, disaster taxonomy, multilingual keyword search

1 Introduction

Online social media services, as Twitter, Facebook or Flickr, have changed the way of communication within communities and groups or between individuals. Monitoring and analysing this continuous flow of user-generated content can yield valuable information, which is not available from traditional sources. Twitters short messages (tweets) can be seen as a dynamic source of information enabling individuals, corporations and government organizations to stay informed. For instance, people are interested in getting advice, opinions, facts, or updates on news or events. Consequently, tweets can give the information to answer the usual 4W questions in the disaster management domain. *What* is happening, *where* and *when* an event is happening and *who* is involved. Within the 140 characters of a tweet the question about the *what* can be answered. The remaining information about the *where*, *when* and *who* need to be extracted from the tweet's metadata. The sender gives the *who-information*, the time stamp gives the *when-information* and the geo-reference gives the *where-information*.

This paper presents a prototype, which will monitor the Twitter stream and detect and analyse diverse kinds of natural disaster events.

The Twitter stream, however, also contains large amounts of meaningless messages, polluted content, spelling or grammatical errors, improper sentence structures and mixed languages, which negatively affect the detection process. To handle these effects a considerable amount of literature has been published on detection approaches. The authors classified the representative techniques for a short review in three categories according to the *event type*, the *detection task* and the *detection method*.

Depending on the event type, the techniques are classified into unspecified and specified event detection. Unspecified events of interest are typically driven by topics, that attract the attention of a large number of users. Because no event information is available numerous features that occur frequently are typically used to detect unknown events (cf. [9, 10]). In contrast, specified event detection aims on known or planned event types. These events could be specified by the related information such as location, time, or performers. The techniques attempt to exploit Twitter's textual content using a wide range of machine learning, data mining, and text analysis techniques (cf. [4, 8]).

According to the detection task, the techniques are classified into new event detection (NED) and retrospective event detection (RED) techniques. Most research is focused on NED, which involve continuous monitoring of signals to exploit the timely information provided by Twitter streams. Knowledge about the event is integrated into the detection, by using filtering techniques as [8] or using additional features such as the location (cf. [5]). RED techniques are more focused on chronological data. The search capabilities allow retrieving individual tweets in response to a query. Because relevant messages may not contain any query term and new shortcuts as hash tags may merge over time, the challenge is identifying relevant messages. So, event retrieval from Twitter data is often focused on temporal and dynamic query expansion techniques (cf. [6]).

Event detection from Twitter draws on different detection methods, including machine learning, data mining, natural language processing, information extraction and information retrieval. The major directions of the approaches are subdivided into *supervised*, *unsupervised* and *hybrid* approaches. Several supervised classification algorithms have been proposed for specified events, including for instance naive Bayes [9], support vector machines [8] or gradient boosted decision trees [7]. Most techniques for unspecified event detection from Twitter streams rely on clustering approaches as expectation–maximization algorithm [1] or threshold-based approaches as [9]. Above that, there are hybrid detection approaches proposed to identify Twitter

messages. In [2], a factor graph model is used, which simultaneously extracts attributes of the event using a supervised conditional random field classifier. The review showed that the detection approaches are very specific regarding their respective aims. Thus they cannot directly be applied in a generic manner.

The detection technique of this paper is a training-based, statistical robust NED algorithm for a specific domain of events. In contrast to the introduced approaches, the prototype described here includes spatial-thematic clustering, temporal monitoring and classifies event types in multiple languages to meet the requirements of the disaster domain (cf. Section 3). The experimental results (cf. Section 4) show the system's performance based on various real-world events.

2 General Framework

This section will detail the data resource Twitter and the programmatically important characteristics of its API (Application Programming Interface). Additionally, the areas of investigation which are monitored by the system will be described

2.1 Data ressource

The developed application exploits Twitter as extensive data resource. Twitter provides real-time access to its worldwide ongoing traffic, called *Firehose*, through its Streaming API. However, only about 1% of all current tweets can be crawled for free.

Since a main focus of the analysis of an event is its location, the application only uses geo-referenced tweets. To send a geo-referenced tweet, the user needs to explicitly allow the Twitter client on his device to access the device's locational sensor (e.g. GNSS sensor). This results in approximately only 2% of all tweets worldwide being geo-referenced.

Besides the capability for keyword filtering, the API also allows for geospatial filtering in terms of bounding boxes. However, these methods are not applicable simultaneously. Nevertheless, these bounding boxes facilitate avoiding the 1% limit.

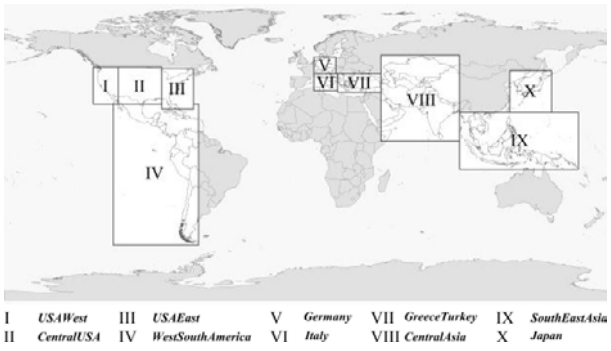
2.2 Investigation Area

The system is able to monitor any area around the globe for potential natural disasters, given a certain training period (c.f, Section 3.3). The areas are limited by a bounding box, due to Twitter's Streaming API constraints (cf. Section 2.1). Figure 1 shows the boundaries of the monitored test areas on a world map.

The areas were selected based on their potential risk of natural disasters such as earthquakes, volcanic eruptions, tornados, etc., and the popularity of Twitter in the countries they contain or overlap. For example, the main part of China is not included, as the Twitter service is blocked there by the Chinese government since 2009.

For example, the bounding box *WestSouthAmerica* (cf. Figure 1) was chosen because of its high risk of volcanic eruptions. The geographical extent ranges from 57° south to 27° north and from 115° west to 64° west.

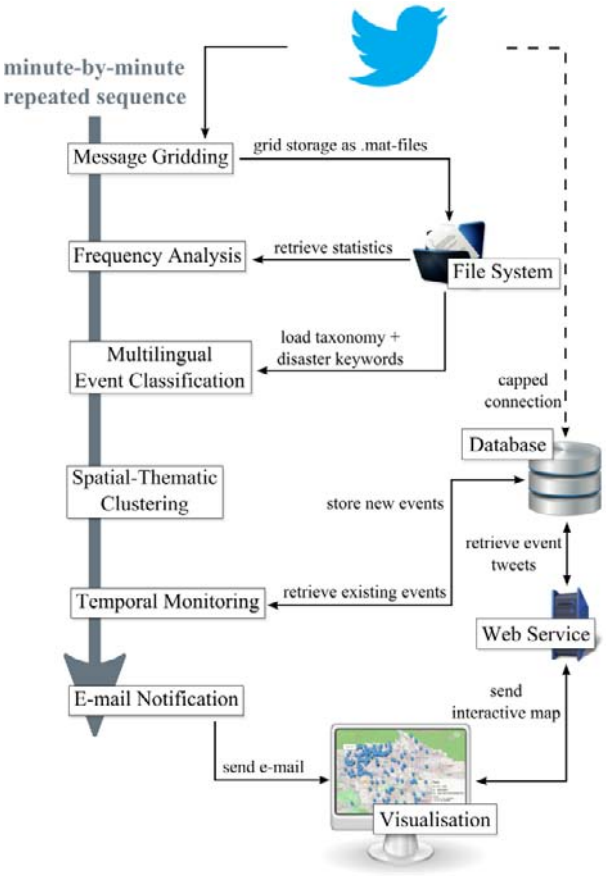
Figure 1: Map of the investigation areas and their names in the system



3 Analysis Workflow

This section will describe the implemented prototype with its analysis workflow in detail. The prototype is an automatic system for multilingual, real-time detection, classification, spatial-thematic clustering and temporal monitoring of natural disaster events. It is not a RED implementation, but a successfully running system that operates twenty-four-seven.

Figure 2: Prototype architecture and analysis workflow



3.1 Prototype Architecture

The implementation is primarily based on the Java programming language with embedded calls to Matlab scripts which e.g. execute numerical calculations. However, the main benefit of using Matlab is the efficient storage of sparse matrices as .mat-files (cf. Section 3.2). The visualization component is based on JavaScript.

Figure 2 depicts the complete architecture and analysis workflow of the implemented prototype exemplified for a single area of investigation. The basic requirement is a broadband internet connection to assure the access to Twitter's *Firehose*. However, the system is tolerant of communication disruptions to the Twitter service, as it immediately tries to re-establish the connection, usually successfully within a few seconds. Thus, even general network failures only result in the system not being able to detect events during that period. Shortly after the internet connection is restored the system will continue unaffectedly.

MongoDB a document-oriented, open-source database is used as data storage technology. As many other NoSQL technologies, it has the advantages of a dynamic schema. MongoDB uses the BSON (binary JSON) format, which is an enhanced version of the JSON format used by Twitter to provide their tweets via the API. Thus, the incoming tweets are stored *on-the-fly* without any further processing needed. Furthermore, the spatial indexing capability of the database allows for extreme fast retrieval of stored tweets based on their location data, i.e. their geographical coordinates.

The incoming tweets are stored in a collection with a capped connection, i.e. after a certain time frame (here 10 minutes) a tweet is deleted from the database. In contrast, the analysed events and their corresponding tweets are stored persistently.

3.2 Message Gridding

The foundation for the powerful detection mechanism is the mapping of the messages to a regular (numerically) 2-dimensional grid based on their geographical coordinates. Thus, also small-scale or regional natural disasters can efficiently be detected and do not disappear in the noise of the tweet baseline of the complete bounding box.

The chosen spacing of the grid points of 0.25° in the test areas is a balance between the speed of detection and the spatial granularity. Depending on the geographical latitude the spacing corresponds to an approximate distance of 25 to 28 km. Consequently, the cells of the grid are 0.25° × 0.25° and exactly cover the area of the respective bounding box. The incoming messages are assigned to the cell that covers the location where they were sent from. The temporal resolution of the system for the test areas is set to one minute, i.e. during one minute the messages are counted per cell and at the end of each minute the grid is stored as .mat-file with an according timestamp (date and time). In general, this time range ensures a sufficient number of tweets w.r.t. the chosen grid spacing to enable a robust statistical evaluation.

3.3 Training Phase

For a statistically robust and reliable detection, a training phase was conducted for each of the ten areas of investigation. In [3] is shown that the usual baseline of tweets of a specific region significantly differs between at least two types of days. This difference is strongly correlated with the percentage of Twitter users in this region who have to work on the next day. The highest accordance across all days of the week lies between 4pm and 6pm local time. To Account for these findings, a 24 hour period (i.e. a day) starts at 5 pm local time respectively. Moreover, the prototype distinguishes between 24 hour periods starting on a Friday, on a Saturday or on another day of the week. With the temporal resolution of one minute, the system stores 1440 grids per day and bounding box.

The training comprised 30 complete 24 hour periods for each of the three types of days. Thus, the mean value and the standard deviation of the amount of tweets could be derived for each cell for every minute of a day (and of course for each bounding box).

3.4 Frequency Analysis

The first indicator of an event in general, is an exceptional increase or decrease of the volume of tweets in a specific region. So far, the prototype only handles the case of increasing Twitter traffic, as it facilitates a meaningful and robust content analysis.

To decide whether an unusual high amount of tweets occurred in a cell during the time step of one minute, the counted number of tweets x is subject to a hypothesis test. Herein, x is checked against the mean value m and the standard deviation s of the cell in the preceding minute. The derived statistical values (m and s) from the training phase are automatically weekly updated in the running system. The null hypothesis H_0 in the test is defined as *no event happened*. The alternative hypothesis H_A consequently is defined as *an event happened*.

The significance level is set to $\alpha = 5\%$ and the power of the test is $1 - \beta = 90\%$. The test value is calculated as

$$k = \frac{x - m}{s} \sim N(0,1) \quad (1)$$

From the critical value for H_0

$$c_{H_0} = N(0,1)_{1-\alpha/2} \quad (2)$$

and H_A

$$c_{H_A} = N(0,1)_{1-\beta} \quad (3)$$

the decentralization parameter

$$\delta = c_{H_0} + c_{H_A} \quad (4)$$

is derived. Finally, if $k > \delta$, the systems accepts the alternative hypothesis H_A and thus identifies a significant rise in tweets in the respective cell.

3.5 Multilingual Event Classification

The next step analyses the content of the tweets in the cells, which were identified through the hypothesis test, to try to assign them a certain class or type of natural disasters.

Therefore, the system retrieves the tweets from the preceding minute and the respective cell from the database based on their timestamp and geographical coordinates. After that, the textual content of the tweets is scanned by a multilingual keyword search for terms related to natural disasters. The 133 terms, which are based on past event experience, were compiled in English. General disaster related terms are also included at this stage to provide a more comprehensive situational awareness for disaster managers.

For each of the 43 languages possibly occurring in the investigation areas (cf. Table 1), the complete list was translated with the aid of the MyMemory REST API[1] and Google Translate[2]. To assure real-time performance, each bounding box was assigned only the languages that are common in its geographic region plus English. For example, the languages for the *WestSouthAmerica* bounding box are Spanish, Portuguese and English, i.e. the German word *Erdbeben* (Eng. earthquake), would not be detected in this bounding box.

Table 1: List of the 43 languages in which the system can identify terms related to natural disasters

Arabian	Spanish	Japanese	Dutch	Tagalog
Azerbaijanian	Persian	Georgian	Polish	Telugu
Bulgarian	French	Khmer	Portuguese	Thai
Bengal	Hindi	Korean	Romanian	Tamil
Bosnian	Croatian	Laotian	Russian	Turkish
Cebuano	Hungarian	Macedonian	Slovak	Urdu
Czech	Armenian	Marathi	Slovenian	Vietnamese
German	Indonesian	Malaysian	Albanian	
Greek	Italian	Maltese	Serbian	

The occurrences of the identified keywords in the retrieved cell are translated back to English and added up term-wise. In result, this yields a list of disaster-related English terms with their respective absolute frequency (cf. Table 2).

For the classification of the event, a hierarchical tree structure (taxonomy) of natural disaster types was established. In this structure, the leaves represent disaster types, which are usually not further distinguished by non-experts in natural language. Each of these leaves is assigned a bag-of-words (BoW) that is virtually unambiguous for the specific event type (cf. Figure 3). Similar to the 133 general disaster terms, the BoWs are derived from investigations of tweets from past events. The union of all BoWs of the child nodes represent the BoW for the respective parent node, e.g. the BoW for the type *Hydrological* is the union of the BoWs of *Flood* and *Tsunami*.

Table 2: Example result of a cell for the multilingual keyword search

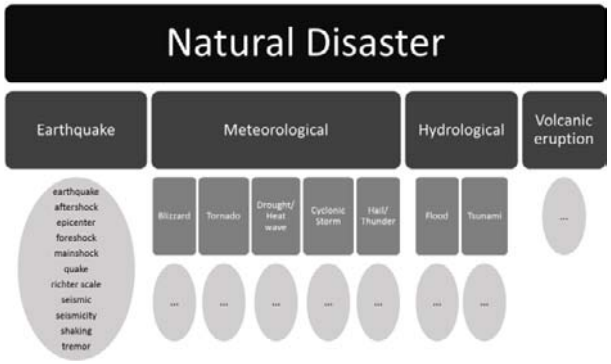
term	**count**
earthquake	10
shaking	4
quake	2
Thunder	1

[1] MyMemory: http://mymemory.translated.net/doc/spec.php
[2] Google Translate: http://translate.google.com

The system starts at the topmost level of the taxonomy and calculates for each node the classification score csc, i.e. the ratio of the number of identified terms that belong to the BoW of the node, and the total number of tweets in the respective cell and minute. A threshold of at least 0.3 is set to assure the relevance of the identified keywords in the current Twitter content. Assuming 20 tweets occurred in the cell in the last minute and the system yielded the results depicted in Table 2, the ratio for *Earthquake* would be 0.8 $((10 + 4 + 2)/20)$ and 0.05 $(1/20)$ for *Meteorological* (the term "thunder" is in the BoW of *Hail/Thunder*).

Only the child node with the highest value is further analysed in an analogous manner. In case of two or more equal values as well as if all child nodes fail to reach the threshold, the parent node is set as type of the event. For example, if the system decided for *Hydrological* in the preceding level, but cannot distinguish between *Flood* and *Tsunami* based on the identified keywords, it will classify the event as *Hydrological*.

Figure 3: Disaster taxonomy with bag-of-words for the natural disaster type *Earthquake*



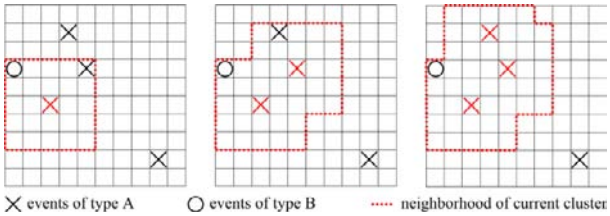
3.6 Spatial-thematic Clustering

For large scale natural disasters that exceed the area of a single cell, the system performs a spatial clustering to aggregate cells that represent the same natural disaster.

The system conducts an algorithm loosely based on the idea of the region growing method in image processing. Here, the initial seed points are the detected and classified single event cells in a time step. In contrast, to image processing, not all cells will be evaluated but only the set of seed points. Hence, in a 24-neighborhood around the seed points the systems searches for others of the same natural disaster type. The rather large neighbourhood tries to account for the inhomogeneous population distribution, which plays a major role in the detectability of events in a specific cell. Thus, even a high impact natural disaster can lead to a diffused detection of affected cells. Figure 4 shows an abstract, exemplary clustering process and the 24-neighborhood of a cell.

The information of the cells in a cluster is fused and from now on interpreted as one single event. Clusters with only one event cell are also referred to as clusters in the following.

Figure 4: Spatial clustering (abstract example); Red colour denotes events of the same spatial-thematic cluster



3.7 Temporal Monitoring

The temporal monitoring operates on the results yielded by the spatial-thematic clustering, i.e. it attempts to link the clusters from preceding time steps to the currently detected ones that refer to the same event.

Therefore, similar to the clustering process, the systems scans the database for detected events that were assigned to the same type of natural disasters and are falling in the merged neighbourhood of the current cluster. In case of a successful search, the associated tweets of the current cluster are persistently stored in relation to the ID of the existing event in the database.

In contrast, if the systems cannot link any existing event in the database to the current cluster, the cluster will be stored with its aggregated information and tweets as a new natural disaster with a unique event ID.

3.8 Notification and Visualisation

After the detection of such a new natural disaster, the system sends an automatic e-mail alert to a given address. The message contains the most important information of the event, such as date, time, geographic place and coordinates, the type of the natural disaster with its csc, the identified disaster keywords and their occurrences as well as the statistical values of the frequency analysis. These include the statistical probability p that the test decided correctly in favour of the alternative hypothesis, the number of tweets in the cluster and its corresponding mean and standard deviation for the minute of the day. Figure 5 shows the e-mail alert for an earthquake of magnitude 3.1 near Los Angeles.

The place is determined through an implemented call to the reverse geocoding service of OpenStreetMap (OSM). The geographical coordinates used as input parameters for the service are the longitude and latitude of the centroid of all tweets in the cluster.

At the end of the notification e-mail, a link to a *node.js* based web service is provided. By clicking on the link, a browser tab opens and visualises all tweets that were assigned to the specific disaster on a map. The basic map data also comes from OSM and is implemented with *leaflet.js*, an open-source JavaScript library for interactive maps.

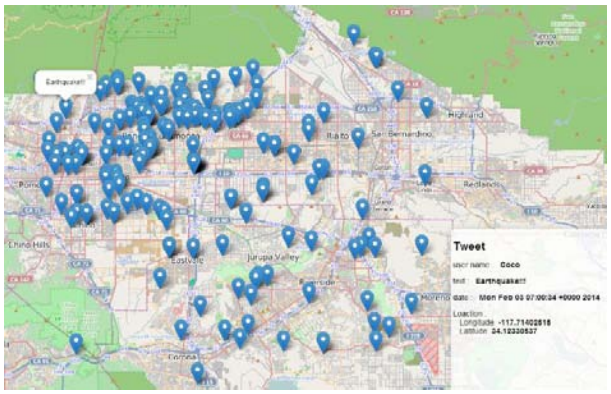
The web service communicates with the database to retrieve the corresponding tweets based on the unique event ID of the natural disaster depicted in Figure 5. The user can view the tweet information by clicking on the tweet markers. The screenshot in Figure 6 depicts the web service response for the earthquake in Figure 5.

Figure 5: E-mail alert with main information of the detected natural disaster

```
TWITTER EVENT !!!

Event ID:   52ef3cb5e1b0f738a1720e0e
Date:       3.2.2014
Time:       7:1:0 UTC
Place:      Fontana, California, United States of America
            Longitude     -117.47142°
            Latitude      34.04352°
Topic:      EARTHQUAKE    0.461
Keywords:
            earthquake    42.0
            quake         2.0
            thunder       2.0
            shaking       1.0
            magnitude     1.0
Statistics:
            %prob         99.88
            #number       99
            mean          7.8061
            std           4.1067
Tweet Map
```

Figure 6: Screenshot of the JavaScript based web service to visualize the event's tweets



Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © CloudMade

4 Experimental Results

Since 01/01/2014 the system automatically detected and analysed a total of 186 natural disasters of varying impact and different types. Table 3 shows the distribution on the different disaster types in absolute numbers and percentage. Due to the lack of a universal definition what constitutes a disaster, an evaluation based on a confusion matrix would not yield meaningful results. Therefore, absolute detection rates are not provided. However, since the actual aim of the system is to analyse natural disasters with impact on people, there is no information loss.

Depending on the disaster type's temporal characteristic the e-mail alert was sent within 40 seconds to 44 minutes from the beginning of the event. As expected, earthquakes are best suited for a detection in real-time (mostly below 2 minutes), because they have an exact starting time and occur unexpectedly. The 4.5 earthquake in Fontana, California on 15/01/2014 at 9:35:19 UTC caused an e-mail alert only 49 seconds later. The event had a statistical probability p of

96.8% and a classification score csc of 0.64. The system stored a total of 1791 tweets that are directly linked to the event containing mentions of "earthquake" (708), "quake" (66), "shaking" (42), etc. The last cluster detection that could be assigned to the event occurred at 10:35 UTC.

Table 3: Absolute number and percentage of automatically detected and analysed disaster types since 01/01/2014

type	number	percentage
Earthquake	78	41.9%
Hail/Thunder	43	23.1%
Natural disaster	18	9.7%
Meteorological	16	8.6%
Tornado	13	7.0%
Volcanic eruption	5	2.7%
Flood	5	2.7%
Tsunami	4	2.2%
Blizzard	2	1.1%
Drought/Heat wave	1	0.5%
Cyclonic storm	1	0.5%
Hydrological	0	0%

Volcanic eruptions have similar characteristics. However, they are rarely located very close to populated places and therefore their effects usually take longer to be noticed by the public. For example, the eruption of the Mt Kelud with its massive emission of ash in Java, Indonesia on 13/02/2014 at 15:50 UTC, was detected at 16:34 UTC.

Other disaster types such as *Flood* or *Hail/Thunder* have no discrete starting but evolve with time. Nonetheless, the system automatically analysed several such events. For example, the flooding caused by heavy rains, in parts of Jakarta, Indonesia on 29/01/2014 in the morning, was detected at 4:45 local time ($p = 99.9\%$ and $csc = 0.63$) and provided 288 tweets with on-site information (mentions: "flood(s)/ing" 84, "inundation" 28, etc.). The severe thunderstorm that hit New York City, USA in the evening of 13. February 2014 was detected at 20:45 local time and 786 tweets could be assigned to the event ($p = 99.4\%$ and $csc = 0.37$) with mentions of "thunder" (221), "lightning" (86), etc.

5 Outlook

The application, introduced in this paper, detects various natural disaster events such as earthquakes, floods or volcanic eruptions (cf. Section 4). The characteristics of the events, e.g. earthquakes and floods, are fundamentally different. On the one hand there is an abrupt punctual event and on the other hand there is a continuous and areal event, but both types are detected and analysed by the prototype. For a comprehensive and reliable evaluation, the approach has to be tested with every possible event type from the taxonomy (cf. Fig. 3). Additionally, the authors will expand the detectable event types on man-made disasters such as industrial accidents, building collapse or smog. Therefore, the hierarchical structure of the taxonomy has to be extended and the appropriate BoWs need to be compiled.

In a next step, the application will be incorporated into a framework for exploring all disaster related information from tweets. The long-term aim is to interpret the tweet's textual content in real time based on machine learning techniques to extract and classify relevant information. This information will help to improve the situational awareness for crisis management.

References

[1] Aggarwal, C., Zhai, C.: A Survey of Text Clustering Algorithms. In: Aggarwal, C.C., Zhai, C. (eds.) *Mining Text Data*, pages 77-128. Springer US 2012.

[2] Benson, E., Haghighi, A., Barzilay, R.: Event discovery in social media feeds. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Association for Computational Linguistics, 2011.

[3] Dittrich, A., Lucas, C.: A step towards real-time detection and localization of disaster events based on tweets. In: *Proceedings of the 10th International ISCRAM Conference*, 2013.

[4] Gu, H., Xie, X., Lv, Q., Ruan, Y., Shang, L.: ETree: Effective and Efficient Event Modeling for Real-Time Online Social Media Networks. In: *Web Intelligence and Intelligent Agent Technology*, IEEE Computer Society, 2011.

[5] Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, ACM, 2010.

[6] Metzler, D., Cai, C., Hovy, E.: Structured event retrieval over microblog archives. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2382138 2012.

[7] Popescu, A.-M., Pennacchiotti, M.: Detecting controversial events from twitter. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, 2010.

[8] Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th international conference on World wide web*, ACM, 2010.

[9] Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: TwitterStand: news in tweets. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, 1653781 2009.

[10] Walther, M., Kaisser, M.: Geo-spatial Event Detection in the Twitter Stream. In: Serdyukov, P., Braslavski, P., Kuznetsov, S., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) *Advances in Information Retrieval*, vol. 7814, pages 356-367. Springer Berlin Heidelberg 2013.

Session:
Theory and Practice

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

A Geometric Configuration Ontology to Support Spatial Querying

Kristin Stock
Nottingham Geospatial Institute
University of Nottingham
Nottingham, United Kingdom NG7 2RD
kristin.stock@nottingham.ac.uk

Abstract

A number of ontologies of spatial relations have been defined in the literature, but most of these are either confined to a small subset of relations, or focussed on language expressions, and not specified geometrically. This paper presents an ontology of geometric configurations, to reflect and specify the range of spatial relations that have been discussed by previous researchers and that are commonly expressed in natural language, and to provide a sufficiently specific definition of the relations to allow them to be executed as spatial queries. Although this work was motivated by a goal to translate natural language describing location into spatial queries, we anticipate wider applications of the ontology for other purposes.

We define a three level ontology, informed by the literature and the study of a corpus of expressions of natural language geospatial location descriptions, and present the concepts and the definition using spatial queries.

Keywords: ontologies, spatial querying, natural language, ontologies.

1 Introduction

There is a distinction between spatial relations as they are used in language; spatial relations as they are described in the qualitative spatial reasoning (QSR) literature, which in some ways attempts to emulate the way spatial relations are used in language, and spatial relation queries that Geographic Information Systems (GIS) or standards-based Spatial Data Infrastructures (SDI) are capable of executing. In GIS and SDI, a restricted range of spatial operators is available, and the only qualitative spatial relations that are currently commonly supported are the basic topological spatial relations and simple buffer/distance calculations [31]. While topology is acknowledged as an important way of describing relations between objects in space, there are a number of other types of spatial relations that are commonly used in natural language, the meaning of which have been explored in detail in both linguistics and QSR. In order to allow geospatial systems to take advantage of the significant work in both spatial linguistics and QSR, it is necessary to develop a mechanism for translating non-topological spatial relations into actual spatial queries that can be executed in a metric system.

To this end, we present an ontology of geometric configurations (GCO). Notionally, it includes parameters to describe (1) spatial relations between pairs of two dimensional objects (for example, topology, orientation, proximity), and (2) the extensions of spatial objects (for example, shape and size). However, in this first version of the GCO, we do not address extension, but focus on spatial relations, and provide a placeholder for extension parameters to be added later. We consider that the combination of relations and extension is required to reflect many of the configurations between geographic objects that are described in natural language, which is the original motivation for this work.

The ontology describes the parameters diagrammatically and specifies them by providing the spatial query that can be used to execute the spatial relation in a GIS or SDI. In some cases, this is straightforward (for example, with topology), but in others, requires more manipulation to convert essentially qualitative parameters into quantitative queries. The ontology presented here brings together much of the QSR work, and specifies methods for converting it into quantitative queries. In many cases, we adopt existing approaches to do this (for example, methods for the quantification of the qualitative notion of proximity have been developed already), while in others, we define a new method.

The GCO provides approaches that could be used in a GIS or SDI, in which data sets may be modelled using points, lines, polygons and complex geometries. However, our work does not extend to 3 dimensional geometries. Finally, for this version, we confine our attention to binary relations.

2 Related Work

A number of linguistically motivated typologies and ontologies have been developed, with the goal of describing a range of spatial relations in terms of their linguistic representations. These are usually focussed around prepositions and explore spatial relations from a linguistic perspective, but do not provide spatially explicit, computational semantics for the terms included, many of which can encompass more than one spatial sense. For example, Coventry and Garrod's [10] typology of relational prepositions includes *in* and *on*, both of which have multiple possible spatial interpretations. Zwarts [35] provides another such typology, based on telicity, and the algebraic properties of different spatial relations. GUM-Space is a very detailed, linguistically motivated spatial ontology based on the General Upper Model (GUM), a task and domain independent linguistically motivated ontology [21]. GUM-Space includes a range of concepts that describe the pertinent content from a natural language spatial expression, but they do not specify a

precise, concrete interpretation, and mappings are required [22].

At the other end of the scale from these linguistically motivated schemes, are schemes that focus on the mathematical interpretation of spatial relations. However, these are usual partial in addressing one or two parameters, with a particular focus on topology. For example, the Ordnance Survey Spatial Relations Ontology[1] includes topological operators, in addition to properties for describing metric location (easting and northing), while the NeoGeo spatial ontology[2] is restricted to topological relations.

Some other typologies and ontologies occupy positions in between these two extremes, including SUMO and OpenCyc[3], both of which provide partial specification, but full, executable semantics are not given [2, 14]. Ontologies that are combined with particular applications techniques include Bucher et al [5], who separate a geometric level ontology from an application level in their topology of spatial relations, and Bitters [4] who proposes to assign weighted probabilities to each relation in his ontology for a given pair of geographic features.

More general typologies are provided by Habel and Eschenback [19], who devise a three dimensional classification of spatial concepts, and Egenhofer and Franzosa [13], who divide spatial relations into topological, metric and ordered relations. In both these cases, full semantics are not provided. Upper level ontologies like DOLCE [16] and BFO's SNAP and SPAN [18] provide foundational concepts for the description of spatial concepts, but do not provide the level of detail required here. Finally, Kemmerer [24] highlights the cross-linguistic differences in spatial relations.

Figure 1: Geospatial Ontology Representation Layers

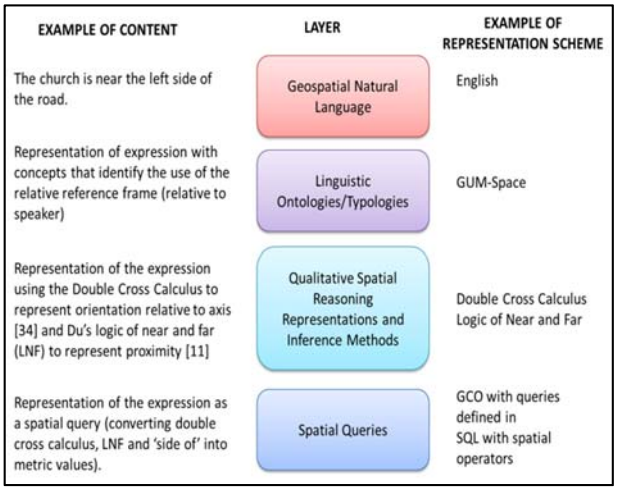


Figure 1 depicts the distinction between the different layers of knowledge representation that are exhibited by the previous work, and the role that the current work plays relative to it. Some of the ontologies described herein cross more than one layer, and to date, the bottom layer has not been separated from the layers above. In some cases the spatial query that is formulated is dependent on the representation used in the QSR layer, but the actual geometric configuration is independent of the QSR representation in our scheme. Thus alternative spatial queries may be developed for different QSR methods. Bennett's [3] work is closely related to the work described herein, and we draw on this work where possible.

3 The Ontology

3.1 Goals

The creation of the GCO was driven by three goals:

1. To create a simple ontology that could express the most common geospatial natural language expressions, rather than every possible permutation of geospatial relations.
2. To focus on the requirements of **geospatial** information. Many of the existing ontologies and typologies include levels of detail that are rarely relevant in the geospatial context. For example, OpenCyc includes spatial relations *hangs from* and *suspended in liquid*, which are not commonly used with geospatial objects.
3. To create concrete, rather than abstract, ontology concepts and properties, that could be specified using a geospatial query that can be executed in a GIS or SDI.

3.2 Three Level Structure

The ontology has three levels. The top level is a division between binary relational parameters (describing the geometric relation between pairs of geometries) and extensional parameters (describing the geometric extension of a single geometry). This accords roughly with the literature, which identifies the importance of relational parameters like topology, orientation and distance [7, 20, 26, 27], and extensional parameters like size and shape [6]. Although the top level of the ontology consists of two branches, the remainder of the work presented here covers the relational parameters only.

The second level of the ontology consists of a series of parameters, being characteristics that may be used to describe the relation or geometry. These include the most commonly discussed relational parameters in the literature (like topology, distance and orientation), as well as other parameters that are relevant for the colloquial description of spatial location. Many of the parameters that have been studied in detail in QSR are accommodated.

The third level of the ontology contains parameter values. In many cases, these are derived from existing literature in QSR and related areas, but they are also considered in terms of the range of ways in a which a parameter may be described, whether qualitative or quantitative. Although the parameter values are given simple language labels for convenience, it must be stressed that the ontology does not aim to describe linguistic spatial relations directly, as have many other ontologies and typologies (for example, SUMO and OpenCyc). We do not describe spatial relation words or phrases, but actual spatial relations that may be encountered in the world. Rather than approaching the problem from the language point of view, we approach from the geometric configuration point of view. The reason for this is that there are multiple ways of describing spatial relations in language

[1] http://data.ordnancesurvey.co.uk/ontology/spatialrelations/
[2] http://socop.oor.net/ontologies/1021
[3] http://www.cyc.com/platform/opencyc

(often the same spatial relation may be described in many different ways), and they differ depending on the language concerned and the context. In this way, multiple language constructs (individual words or more complex phrases) may be mapped to the same geometric configuration in the ontology. It is also likely that in some cases, language descriptions will map to multiple parameter values, which must all be true in order for the natural language expression to be fully realised. Finally, language based ontologies are often much more complex than this ontology, because they encompass the myriad different ways of describing a spatial relation. We aim to maintain the simple three level structure for this ontology, as much of the complexity is in the way language expresses relations, rather than the relations themselves.

The spatial relations are defined using spatial queries, which are provided using standard SQL syntax, according to ISO 13249-3 [23]. This standard was used because the goal of the work is to map geometric concepts to spatial queries, and ISO 13249-3 is widely used (sometimes with minor syntactic variations) by most GIS systems, along with relevant SDI standards. The queries included are designed to accommodate point, line and polygon geometries. We use previous work in the QSR literature to define these in some cases, as follows, but other schemes could easily be substituted to suit the required purposes.

Topology: We adopt a simple set of five parameter values, including the touches spatial relation (excluded from RCC5), but excluding the distinction between tangential and non-tangential proper parts [8,9]. This is because we have not found any evidence that the distinction is commonly made in natural language descriptions of spatial relations. We also exclude inverse relations (contains in addition to within), as the same relation can be expressed by reversing the geographic features concerned.

Distance: We adopt the simple logic of near and far (LNF) of Du et al [11], but other more complex schemes (and particularly, more advanced methods for calculating nearness) could be substituted [6, 15, 33]. LNF determines nearness using buffer zones and a fixed sigma value that is selected manually for the activity concerned, and the queries we define reflect this. We also include a quantitative representation of distance that is commonly encountered in natural language. It involves description of distance using a simple quantity and unit, the latter being either spatial (for example, 300 metres, 1 mile) or temporal, in which case it is includes an explicit or implicit mode of travel (for example, 3 minutes' walk, 5 hours' drive).

Linear Orientation: We adopt Dugat et al's [12] set of orientations, as it is the most comprehensive set, of several similar alternatives [1, 6, 26].

Horizontal Projective Orientation: We adopt the simple scheme of Clementini et al [6], defining left/right and back/front as semi-circles, rather than quadrants. This is because these simple relations more closely reflect the most commonly used linguistic expressions and because combined expressions like '*left and in front of*' may be determined by combining their individual components (left, in front of). We do not adopt the between relation proposed by more recent work by Clementini et al [7], which is similar in principle to relations provided in the Double Cross Calculus [34] and the Dipole Calculus [26], since these are also not commonly used in natural language (except for examples that include three objects, and are thus ternary relations and out of our scope).

Direction: We adopt Goyal and Egenhofer's [17] 9 cell model for cardinal direction relations, based on the minimum bounding rectangle of the reference object. We chose this basic model due to its simplicity and the absence of empirical evidence that the subsequent extensions more closely reflect natural language.

Adjacency: We refine the Wordnet definition of adjacency used by Klien and Lutz [25] to require that objects must be either disjoint or touching (as overlapping objects are unlikely to be next to each other). This is thought to be only partially expressive of the adjacency notion, which also includes a consideration of intervening objects of the same type, and is also in many cases part of more specific specialisations of adjacency like alongside (both approximately parallel and adjacent). However, these aspects are handled at the level of contextual analysis of a natural language expression, rather than the geometric configuration ontology.

Collocation: The notion of objects being 'in the same place' is commonly encountered in natural language, and described varying degrees of proximity. It is expressed often using the 'on' (*situated on the*) or 'at' (*located at the junction*) and 'in' (*in the area*) prepositions with varying degrees of precision. Previous work in this area is limited, so we adopt a simple scheme based on topological relations.

Object Parthood: A significant amount of work has addressed the notion of object parthood, and the language used to discuss it [29, 30, 32]. This previous work mainly focuses on identifying different types of parthood, depending on the functional relationship between parts and wholes, type similarity and separability. However, language descriptions of parthood commonly address specific parts of an object (start, end, middle), and our interest is in providing a mechanism for defining which part of a whole is referred to by particular specifications. For this purpose, we define a range of different specific parts. Our interest is in words that specify particular parts of wholes that apply across a range of feature types, rather than parthood relations that are specific to a particular feature type (for example, desert-oasis). In addition to the use of both orientation and direction parameter values as spatial relations (*x is north of y*), they may also be used as adjectives to define some part of an object. We treat these as spatial relations between some part of the object and its entirety.

Figure 2 shows the ontology parameters and parameter values, along with axioms and the queries that define them.

4 Examples

The following examples illustrate the mapping from natural language to queries via the GCO, and are based on results from a recently conducted questionnaire (to be reported in a future publication).

1. 'A train station in Nottingham' maps to the *contain* GCO concept, which maps to a simple query using the contains spatial relation, with the geometry for Nottingham, to show the area in which the desired train station might exist.

2. 'The street next to Jasmine Cottage' maps to the *adjacent* GCO concept, which uses the Jasmine Cottage geometry with the simple within distance method to create a query that selects streets in the appropriate area.
3. 'The monument outside the Town Hall' maps to both the *adjacent* and *disjoint* GCO concepts, and would require the two queries to be combined conjunctively so that only areas that are within a specified distance but not touching are included.
4. 'Development along the Trent River network' maps to the *parallel*, *alongside* and *side (part)* GCO concepts. As with the previous example, combined conjunctively.

These last two examples illustrate cases in which query composition is required. The default approach is to combine the queries conjunctively, so that all clauses must be fulfilled to define the area of interest. Future work will explore query composition methods in more detail.

Table 1: Counts for Each Relational Parameter

RELATIONAL PARAMETERS	**Qty**	**%**	**GCO**
path direction	128	**27.2%**	×
collocation	68	**14.4%**	✓
topology	44	**9.3%**	✓
direction	32	**6.8%**	✓
object parthood	32	**6.8%**	✓
distance	30	**6.4%**	✓
adjacency	19	**4.0%**	✓
horizontal projective orientation	17	**3.6%**	✓
traversal	16	**3.4%**	✓
throughness	13	**2.8%**	✓
alignment	11	**2.3%**	✓
joining	11	**2.3%**	Partial
vertical projective orientation	9	**1.9%**	Out of scope (3D)
distribution	9	**1.9%**	Partial
possession	7	**1.5%**	✓(topology)
betweenness	5	**1.1%**	Out of scope (tertiary)
surroundedness	5	**1.1%**	×
splitting	3	0.6%	Out of scope (tertiary)
aroundness	3	0.6%	×
sidedness	2	0.4%	Out of scope (tertiary)
boundedness	2	0.4%	Partial
protrusion	2	0.4%	×
linear orientation	1	0.2%	✓
oppositeness	1	0.2%	Out of scope (tertiary)
trajectory	1	0.2%	Out of scope (3D)
TOTALS	471	100%	

5 Evaluation

By way of partial evaluation of the focus and completeness of the ontology, we now present an analysis of the spatial parameters and parameter values used in a random selection of 200 spatial clauses from the Nottingham Corpus of Geospatial Language (NCGL) [28]. The NCGL is a publicly available[4] corpus containing only expressions that describe spatial location harvested from web sites, and is thus uniquely placed to evaluate the coverage of the GCO, unlike most other corpora that contain a wide selection of non-spatial language as well, making an evaluation of this kind very time-consuming.

In the 200 clauses from the NCGL that were examined, 471 distinct spatial relational parameter values were identified. Table 1 presents the quantities of each parameter value, listing all parameter values that were encountered, whether or not they appeared in the ontology. While 200 is not a very large quantity, the percentages were stable with the addition of the last 50 clauses in terms of the broad colour coded categories that indicate the percentage of spatial location expressions that the parameter in question includes (in the percent column, >10% shown in red; 3-9% inclusive shown in blue; 1-2% inclusive shown in brown).

While path location is the most frequent of the parameters, it was not included in this version of the GCO, because it is a grouping of a number of concepts that describe both geometric configuration and direction of movement, including expressions like *from*, *to*, *onto*, *away from*, *leave*, *upwards* and *uphill*. These deserve special treatment and are therefore beyond the scope of this version of the GCO.

The GCO covers all of the most frequently occurring relational parameters other than path direction, and 66% of the total set of relational parameters found in the 200 expressions. As can be seen from the right-most column in Table 1, future extensions to include tertiary relations and 3D would be useful to accommodate a more complete range of concepts, along with the addition of path direction and the development of the extensional parameters branch of the GCO.

6 Conclusions

This paper has presented a Geometric Configurations Ontology that is designed to support natural language querying, but has wider applications, and could be implemented as an interface to make a range of new spatial query operators available through standard GIS and SDIs. In future work we are conducted more extensive evaluations with empirical studies, and plan to extend the scope of the current ontology.

[4] http://geospatiallanguage.nottingham.ac.uk/

Acknowledgements

The work described in this paper was funded by Ordnance Survey under the NaturalGeo Project.

References

[1] M. Aurnague and L. Vieu. Towards a Formal Representation of Space in Language: A Commonsense Reasoning Approach. *IJCAI-93 Workshop on Spatial and Temporal Reasoning*, pages 123-158, 1993.

[2] J. Bateman and S. Farrar. Towards a generic foundation for spatial ontology. In Achille C. Varzi and Laure Vieu. *Formal Ontology in Information Systems: Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS-2004)*. Amsterdam, pages 237-248, 2004.

[3] B. Bennett. Spatial Relations in the AURA Knowledge Base. Version 1.0. Unpublished Report, School of Computing, University of Leeds, 2012.

[4] B. Bitters. Spatial relationship networks: Network theory applied to GIS data. *Cartography and Geographic Information Science*, 36(1):81-93, 2009.

[5] B. Bucher, G. Falquet, E. Clementini, and M. Sester. Towards a typology of spatial relations and properties for urban applications, Usage, Usability, and Utility of 3D City Models – European COST Action TU0801, 2012.

[6] E. Clementini, P. Di Felice, and D. Hernández. Qualitative Representation of Positional Information, *Artificial Intelligence*, 95, 317-356, 1997.

[7] E. Clementini, S. Skiadopoulos, R. Billen and F. Tarquini. A Reasoning System of Ternary Projective Relations. *IEEE Trans. Knowl. Data Eng*. 22(2): 161-178, 2010.

[8] A. Cohn, B. Bennett, J. Gooday and N. Gotts. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica*, 1, 275-316, 1997.

[9] A. Cohn and S. Hazarika. Qualitative Spatial Representation and Reasoning: An Overview. *Fundamenta Informaticae*, 43: 2-32, 2001.

[10] K.R. Coventry and S.C. Garrod. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press, East Sussex, 2004.

[11] H. Du, N. Alechina, K. Stock, and M. Jackson. The logic of NEAR and FAR. *COSIT 2013: Conference on Spatial Information Theory*, Scarborough, UK, 2013.

[12] V. Dugat, P. Gambarotto and Y. Larvor. Qualitative Theory of Shape and Orientation, *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '99)*, pages 45-53, 1999.

[13] M.J. Egenhofer and R. Franzosa. Point-Set Topological Spatial Relations. *International Journal of Geographical Information Systems*, 5: 161-174, 1991.

[14] S. Farrar and J. Bateman. General Ontology Baseline. Collaborative Research Center for Spatial Cognition, University of Bremen, Germany. I1-[Main.OntoSpace]: D1. SFB/TR8 internal report http://www.ontospace.uni-bremen.de/pub/FarrarBateman04-i1-d1.pdf. 2005.

[15] A.U. Frank. Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. *Journal of Visual Languages and Computing*, 3: 343-371, 1992.

[16] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari and L. Schneider. Sweetening Ontologies with DOLCE. In A. Gómez-Pérez, V.R. Benjamins, editors, *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002*, Siguenza, Spain, Springer Verlag, 166-181, 2002.

[17] R. Goyal and M. Egenhofer. The Direction-Relation Matrix: A Representation for Directions Relations between Extended Spatial Objects, *Proc. Univ. Consortium for Geographic Information Systems Science (UCGIS) Ann. Assembly and Summer Retreat*, June 1997.

[18] P. Grenon and B. Smith. SNAP and SPAN: Towards Dynamic Spatial Ontology, *Spatial Cognition and Computation*, 4(1): 69-103, 2004.

[19] C. Habel and C. Eschenbach. Abstract structures in spatial cognition. In C. Freksa, M. Jantzen and R. Valk, editors, *Foundations of Computer Science. Potential – Theory – Cognition*, Springer, Berlin, pages 369–378, 1997.

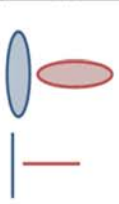
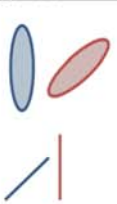
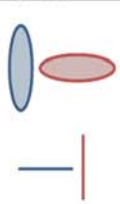
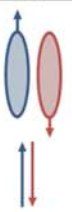
[20] T. Hahmann and M. Grüninger. A naïve theory of dimension for qualitative spatial relations. In *Proc. of the Symposium on Logical Formalizations of Commonsense Reasoning (CommonSense 2011)*, AAAI Spring Symposium, 2011. AAAI Press, 2011.

[21] J. Hois, T. Tenbrink, R. Ross and J. Bateman. GUM-Space: The Generalized Upper Model spatial extension: a linguistically-motivated ontology for the semantics of spatial language Technical Report. Universität Bremen SFB/TR8 Spatial Cognition, http://www.ontospace.uni-bremen.de/ontology/TechnReport09GUMspace.pdf, 2009.

[22] J. Hois and O. Kutz. Natural Language meets Spatial Calculi. *Spatial Cognition VI. Learning, Reasoning, and Talking about Space. 6th International Conference on Spatial Cognition*, pages 266-282, 2008.

[23] International Standards Organisation. Information technology – Database languages – SQL multimedia and application packages, Part 3: Spatial. 13249-3:2011 International Standard, 2011.

Figure 2: The Geometric Configuration Ontology

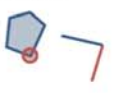
Parameter		Values							
TOPOLOGY (t): Are the objects connected and how?	Label	overlap(a,b)	touch(a,b)	contain(a,b)	disjoint(a,b)	equal(a,b)			
	Illustration								
	Query	ST_Overlaps(a,b) = 1	ST_Touches(a,b) =1	ST_Contains(a,b) = 1	ST_Disjoint(a,b) = 1	ST_Equals(a,b) = 1			
	Axioms								
DISTANCE (ds): How close are the objects to each other?	Label	distance 0 all points(a,b)	distance 0 any point(a,b)	very near(a,b)	near(a,b)	neither near nor far(a,b)	far(a,b)	x spatial units apart (a,b,x)	x temporal units apart by travel at y velocity[1](a,b,x,y)
	Illustration								
	Query (WHERE clause)	ST_Equals(a,b) = 1	ST_Touches(a,b) =1	ST_DWithin(a,b,σ)	ST_DWithin(a,b,2σ)	(ST_Distance(a,b) > 2σ) AND (ST_Distance(a,b) < 4σ)	NOT ST_DWithin(a,b,4σ)	ST_Distance(a,b) = x ST_Distance(ST_Centroid(a), ST_Centroid(b)) = x	ST_Distance(a,b) = xy ST_Distance(ST_Centroid(a), ST_Centroid(b)) = xy
	Axioms	ds.zeroAllPoints(a,b) ≡ t.equal(a,b)	ds.zeroAnyPoint(a,b) ≡ t.touch(a,b)	ds.veryNear(a,b) ≡ (t.disjoint(a,b) ∨ t.touch(a,b))	ds.near ≡ (t.disjoint(a,b) ∨ t.touch(a,b))	ds.neitherNearNorFar(a,b) ≡ t.disjoint(a,b)	ds.far(a,b) ≡ t.disjoint(a,b)	ds.spatialUnitsApart(a,b) ≡ t.disjoint(a,b)	ds.temporalUnitsApart(a,b) ≡ t.disjoint(a,b)
LINEAR ORIENTATION(lo): How are linear objects oriented relative to each other?	Label	parallel(a,b)	perpendicular(a,b)	diagonal(a,b)	orthogonal(a,b)	antiparallel(a,b)	crossed(a,b)		
	Illustration								
	Query (WHERE clause)	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(a)))) – MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(b)))) = 0 ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = 0	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(a)))) – MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(b)))) IN (π/2, 3π/2) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (π/2, 3π/2)	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(a)))) – MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(b)))) IN (π/4, 3π/4, 5π/4, 7π/4) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (π/4, 3π/4, 5π/4, 7π/4)	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(a)))) – MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(b)))) IN (0, π/2, π, 3π/2) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (0, π/2, π, 3π/2)	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(a)))) – MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(b)))) = π ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = π	MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(a)))) – MaxAzimuth[2](ST_Boundary(SSRectangle[3](ST_ConvexHull(b)))) IN (π/2, 3π/2) AND ST_Overlaps(a,b) ST_Azimuth(ST_StartPoint(a), ST_EndPoint(a)) - (ST_Azimuth(ST_StartPoint(b), ST_EndPoint(b)) = IN (π/2, 3π/2) AND ST_Overlaps(a,b)		
	Axioms	lo.parallel(a,b)⇔lo.orthogonal(a,b)	lo.perpendicular(a,b)⇔lo.orthogonal(a,b)			lo.antiparallel(a,b)⇔lo.orthogonal(a,b)	lo.crosses(a,b)⇔t.overlap(a,b)		
HORIZONTAL PROJECTIVE ORIENTATION(hpo): How are objects oriented to each other relative to a projected axis?	Label	in front of(a,θ[4],b)	behind(a,θ,b)	left(a,θ,b)	right(a,θ,b)	alongside(a,θ,b)			
	Illustration								
	Query	ST_Angle(ST_Azimuth(a,b), θ) < π/2	ST_Angle(ST_Azimuth(a,b), θ) > π/2	(ST_Azimuth(a,b) < θ) AND (ST_Azimuth(a,b) > θ±2 π)	(ST_Azimuth(a,b) > θ) AND (ST_Azimuth(a,b) < θ±2 π)	ST_Angle(ST_Azimuth(a,b), θ) IN (π, 3π/2)			
	Axioms								

[1] Calculated using a lookup table listing average speed for different modes of travel (walking, driving, etc) to relate to natural language expression. The temporal distance must use the same units as the velocity unit and conversion may be required (e.g. seconds and metres/second).

[2] MaxAzimuth is a user defined function that finds the azimuth of the longest side of the boundary of the smallest surrounding rectangle (by testing ST_Length for the first and second sides only), thus representing the direction of the elongated polygon. A value of zero is returned if the side lengths are the same (and thus the polygon is not elongated).

[3] SSRectangle is a user defined function that implements an algorithm to computer the smallest surrounding rectangle of the convex hull. This differs from the minimum bounding rectangle or envelope (which is available through the ST_Envelope method) in that it is oriented in the direction that makes the smallest rectangle, whereas the envelope is oriented to the x and y axes of the coordinate reference system. The algorithm for calculating the smallest surrounding rectangle is described in http://gis.stackexchange.com/questions/22895/how-to-find-the-minimum-area-rectangle-for-given-points

[4] θ is the azimuth of the direction of the front of a.

Parameter		Values							
DIRECTION (dr): What is the cardinal direction from one object to the other?	**Label**	north(a,b)	south(a,b)	west(a,b)	east(a,b)	northEast(a,b)	northWest(a,b)	southEast(a,b)	southWest(a,b)
	Illustration								
	Query (WHERE clause)	MinY[1](ST_Envelope(b)) >= MaxY(ST_Envelope(a)) AND MinX(ST_Envelope(b)) >= MinX(ST_Envelope(a)) AND MaxX(ST_Envelope(b)) <= MaxX(ST_Envelope(a))	MaxY (ST_Envelope(b)) <= MinY(ST_Envelope(a)) AND MinX(ST_Envelope(b)) >= MinX(ST_Envelope(a)) AND MaxX(ST_Envelope(b)) <= MaxX(ST_Envelope(a))	MaxX (ST_Envelope(b)) <= MinX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MinY(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MaxY(ST_Envelope(a))	MinX (ST_Envelope(b)) >= MaxX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MinY(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MaxY(ST_Envelope(a))	MinX (ST_Envelope(b)) >= MaxX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MaxY(ST_Envelope(a))	MaxX (ST_Envelope(b)) <= MinX(ST_Envelope(a)) AND MinY(ST_Envelope(b)) >= MaxY(ST_Envelope(a))	MinX (ST_Envelope(b)) >= MaxX(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MinY(ST_Envelope(a))	MaxX (ST_Envelope(b)) <= MinX(ST_Envelope(a)) AND MaxY(ST_Envelope(b)) <= MinY(ST_Envelope(a))
	Axioms								
ADJACENCY (a): Are objects adjacent to each other?	**Label**	adjacent(a,b)							
	Illustration								
	Query (WHERE clause)	(ST_DWithin(a,b,σ)) AND ((ST_Touches(a,b) =1) OR (ST_Disjoint(a,b) =1))							
	Axioms	a.adjacent(a,b) ⇒ ds.near							
COLLOCATION (cl): Are objects in the same place?	**Label**	within collocated(a,b)	exactly collocated(a,b)	substantially collocated(a,b)	approximately collocated(a,b)				
	Illustration								
	Query (WHERE clause)	ST_Contains(a,b) = 1	ST_Equals(a,b) = 1	(ST_Overlaps(a,b) = 1) AND (ST_Area(ST_Difference(a,b) > ST_Area(a)/2[2])	ST_DWithin(a,b,σ)				
	Axioms	cl.within collocated (a,b) ≡ t.contain(a,b)	cl.exactly collocated (a,b) ≡ t.equal(a,b)	cl.substantially collocated (a,b) ≡ t.overlap(a,b) cl.exactly collocated (a,b) ⇒ cl.substantially collocated (a,b)	cl.substantially collocated (a,b) cl.exactly collocated (a,b) cl.substantially collocated (a,b) ⇒ cl.approximately collocated (a,b) cl.exactly collocated (a,b) ⇒ cl.approximately collocated (a,b) cl.within collocated (a,b) ⇒ cl.approximately collocated (a,b)				
OBJECT PARTHOOD (op): Which part of the object is of interest?	**Label**	part(a,b)	whole(a,b)	rest(a,b,c)	front(a,b,D)	back(a,b,D)	left side(a,b,D)	right side(a,b,D)	middle (a,b)	corner (a,b, σ)
	Illustration									
	Query (WHERE clause)	ST_Contains(a,b) = 1	ST_Equals(a,b) = 1	ST_Equals(ST_Union(a,b),c) = 1	FrontGeometry[3](a, D) = b	BackGeometry (a, D) = b	LeftSideGeometry (a, D) = b	RightSideGeometry (a, D) = b	ST_Centroid(a) = b	PolygonAngle[4](ST_Intersection(a,b)) < σ AND PolygonAngle(ST_Intersection(a,b)) > 0 (ST_Touches(a,b) = 1) AND ST_Intersection(ST_Boundary(a), ST_Boundary(b) IS NOT NULL AND (ST_Azimuth(a) <> ST_Azimuth(b))
	Axioms	op.part(a,b) ≡ t.contain(a,b)	op.whole(a,b) ≡ t.equal(a,b)	op.rest(a,b,c) ⇒ op.part(a,c) ∧ op.part(a,b) op.rest(a,b,c) ⇒ ~t.overlap (a,b) op.rest(a,b,c) ⇒ t.touch (a,b)	op.front(a,b,D) ⇒ ~ op.back (a,b,D) ∧ ~ op.left side (a,b,D) ∧ ~ op.right side(a,b,D)	op.back(a,b,D) ⇒ ~ op.front(a,b,D) ∧ ~ op.left side (a,b,D) ∧ ~ op.right side(a,b,D)	op.left side(a,b,D) ⇒ ~ op.back (a,b,D) ∧ ~ op.front (a,b,D) ∧ ~ op.right side(a,b,D)	op.right side(a,b,D) ⇒ ~ op.back (a,b,D) ∧ ~ op.front (a,b,D) ∧ ~ op.left side(a,b,D)	op.middle (a,b,D) ⇒ ~ op.back (a,b,D) ∧ ~ op.front (a,b,D) ∧ ~ op.left side(a,b,D) ∧ ~ op.right side(a,b,D)	op.corner (a,b,D) ⇒ ~ op.junction (a,b,D)

[1] MaxX and similar user defined functions provide the maximum X coordinate of the ST_Envelope.

[2] For an overlap of 50% of a with b. Different ratios could be used as required.

[3] FrontGeometry(x, D, σ) is a user defined function that returns the geometry containing all the line segments in the simplified geometry of x whose centroid is closer to the front of the Relative Minimum Bounding Rectangle (RMBR) around x and orthogonal to D than the back, and for whom the angle between the line segment and the front of the RMBR is less than the angle between the line segment and whichever side of the RMBR is closer to the centroid of the line segment. Equivalent definitions for BackGeometry, LeftSideGeometry and RightSideGeometry exist, as per Stock (*).

[4] PolygonAngle(x) is a user defined function that returns the angle at point x of the sides of the polygon of which x is a part.

[24] D. Kemmerer. The semantics of space: integrating linguistic typology and cognitive neuroscience. *Neuropsychologia*, 44, 1607-1621, 2006.

[25] E. Klien and M. Lutz. The role of spatial relations in automating the semantic annotation of geodata. *Conference on Spatial Information Theory*, 2005.

[26] R. Moratz, J. Renz and D. Wolter. Qualitative Spatial Reasoning about Line Segments. *ECAI 2000*, 234-238, 2000.

[27] C. Schlieder. Representing Visible Locations for Qualitative Navigation. In N. Carrete and M. Singh, M.G., editors, *Qualitative Reasoning and Decision Technologies*, 523-532, CIMNE, 1993.

[28] K. Stock, R.C. Pasley, Z. Gardner, P. Brindley, J. Morley, J. and C. Cialone. Creating a corpus of geospatial language. *COSIT 2013: Conference on Spatial Information Theory*, Scarborough, UK. Lecture Notes in Computer Science (LNCS) 8116, 2013.

[29] A.C. Varzi. Parts, Wholes and Part-Whole Relations, The Prospects of mereotopology. Data and Knowledge Engineering 20, 259–286, 1996.

[30] L. Vieu and M. Aurnague. Part-of relations, functionality, dependence. In M. Aurnague, M. Hickmann and L. Vieu, editors, *The categorization of spatial entities in language and cognition*, John Benjamins (Human Cognitive Processing 20), pages 307-336, 2007.

[31] P. Vretanos. Open Geospatial Consortium Filter Encoding Specification. OGC 04-095, 2005.

[32] X. Wood and A. Galton. A Taxonomy of Collective Phenomena. *Applied Ontology*, 4, 267-292, 2009.

[33] M.F. Worboys. Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15(7): 633–651, 2001.

[34] K. Zimmerman and C. Freksa. Qualitative Spatial Reasoning Using Orientation, Distance and Path Knowledge. *Applied Intelligence*, 6:49-58. 1996.

[35] J. Zwarts. Prepositional aspect and the algebra of paths. *Linguistics and Philosophy*, 28, 739-779, 2005.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Spatiotemporal Data Complexity in electronic Airport Layout Plan and its visualization

Shyam Parhi
Airport Engineering Division
AAS-100
Federal Aviation Administration
Washington DC 20591
USA
Email: shyam.parhi@faa.gov

Abstract

Airports GIS is a web portal consisting of a few application modules. It allows authorized users of FAA (Federal Aviation Administration) to submit changes to airport data. One of these applications is electronic Airport Layout Plan (eALP). The main purpose of building this application is to replace the paper copy version of ALP by digital copy. The visualization of the digital copy on the computer screen poses lot of challenges. Spatiotemporal nature of data brings added complexity.

Keywords: Airport, GIS, spatiotemporal, visualization

1 Introduction

Airports GIS can be accessed from internet at airports-gis.faa.gov by authorized users. It consists of several applications and the number of applications is growing every year. One important application discussed here is electronic Airport Layout Plan (eALP). The whole idea of creating such application is to move from paper copy to digital display which brings about lots of challenges that are discussed here. The data is geospatial and temporal and the data needs to be visualized as best as possible and as accurately as possible.

2 electronic Airport Layout Plan

The eALP helps support Next Generation (NextGen) of air transportation which is an FAA wide initiative. Hence it is not surprized that the whole effort needs enterprise level workflow and its implementation at enterprise level as well. Also to safeguard data and retain its integrity, one has to keep in mind some sort of digital signature as we are moving away from paper version.

2.1 Some Basics about eALP

There are approximately over 13,ooo airports and 5800 heliports in USA. Some of their classifications are large hub, small hub, and towered airports. Some are NPIAS (National Plan of Integrated Airport Systems) airports funded by FAA, and others are non NPIAS. NPIAS airports get grants from FAA for different activities on the airport. One of the requirements for these airports to get grants is to prepare and update their Airport Layout plan. However, airport data is changing from time to time. For example, a new runway or taxiway construction requires an update on ALP. It is very difficult to maintain paper copies of ALPs. Since these copies are housed at multiple locations, the data for a specific feature may vary in different locations. This spatiotemporal nature of data can be easily handled through eALP and versioning of eALPs. The data has to be visualized using certain software. Typically we use ESRI product to visualize such data. Some data is needed to be displayed with great accuracy up to under a foot. This requires good visualization techniques and tools.

.

2.2 Dataflow for an Enterprise Model

The dataflow for any enterprise model is crucial. It needs completely different architecture to build an enterprise level application. The simple reason behind such dataflow is that the activities related to specific data are processed at district, region, and headquarters level. Sometimes it has to be coordinated through different lines of business. At times data has to be updated dynamically through web services. All these have to be considered for a good design of enterprise system. Some data are interrelated due to their inherent nature. This needs special attention.

2.3 How data is processed

The Airport data passes through various steps before it is gathered for building eALP. First, the data is entered in the Airports GIS by Airports using a separate Survey module, also available at Airports GIS. The surveyor for a particular Airport uploads data, say for a construction project, to the Airports GIS portal along with some Statement of Work (SOW). This SOW is verified by Airport District Office. The data provider also submits geo-referenced imagery which could be aerial, satellite, or LIDAR along with a plan for these imageries. National Geodetic Survey validates and verifies the

safety critical data. Once the data is verified, it is stored at NASR (National Airspace System Resources) database which is used by all Lines of Business. This verified data is assembled for the purpose of initiating eALP process. This process has many mandatory and a few voluntary steps.

2.4 How data is reviewed

This process is very important. The data is reviewed first at Airport level. Then the system sends one automatic email to Airport District Office to review it. The process is repeated for respective Regions and headquarter. There are some security measures in place so that the data cannot be altered by an unauthorized person. The industry standard convention of using digital signature protects the data. Sometimes the coordination process needs significant discussion with other Lines of Business. This is facilitated by application. There is also some kind of comment board to gather comments and replies at the same time. The coordination process plays significant role as there may arise complete disagreement on some specific feature of eALP among two Lines of Business. Resolving them through traditional methods does not work. The application provides necessary Graphical User Interface to help coordinate the process easily. The delegation process is also implemented. In a vast organization like FAA, one person cannot be designated permanently to perform some task. The delegation process plays a great role in accelerating some tasks and meeting deadlines.

2.5 Technical Challenges

We have lot of technical challenges for implementation of eALP. It is specially complicated when different lines of business own different environments of the same application. For example, one Line of Business develops the application whereas another Line of Business deploys it so that it is accessible to authorized users via internet. Temporality is a big piece in the development of eALP. Data is changing continuously on the airports. How can paper copies keep track of all these changes? When one talks about eALP, one has to realize that temporary storage of data at enterprise level, applied programmatically for all airports, and deletion of some data when they are not used or will not be used in future, pose significant challenge in building eALP. A lot of architectural and design changes are required. Also, we have been using ESRI viewer to visualize the data on the airport. Each software has its own limitation. The integration of any visualization tool with the application brings more challenges.

3. Conclusion

The eALP was recently deployed into production system. However, it is still in the pilot phase. We learned a lot from our pilot phase program. We now know the expectations of airports is that the eALP generated as a pdf file from a process in Airports GIS should look like almost legacy ALP, even if the display on the computer screen can be made better than the look and feel of legacy ALP. That needs lot of work which we plan to complete this year. We will continue to improve the spatiotemporal portion of eALP in next few months. We are also improving ESRI's ArcGIS viewer capability by custom modifying some of its portions. Three dimensional capability of viewing a taxiway or runway or a building is still not available in the existing eALP.

.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Towards Spatio-temporal Data Modeling of Geo-tagged Shipping Information

Amin Mobasheri *
GIScience research group, Heidelberg university
Berliner str.48, 69120
and
Cluster of Excellence Asia & Europe in a Global Context, Voss str. 2, 69115
Heidelberg, Germany
Amin.mobasheri@geog.uni-heidelberg.de

Mohamed Bakillah
GIScience research group, Heidelberg university
Berliner str.48, 69120, Heidelberg, Germany
and
Department of Geomatics Engineering
University of Calgary, Alberta, Canada
Mohamed.Bakillah@geog.uni-heidelberg.de

Abstract

Spatio-temporal data models deal with capturing information characterized by both spatial and temporal semantics. In this paper we review current approaches for spatio-temporal data modelling and present out initial results for selecting the most relevant approach: Object-Oriented modeling for means of modeling geo-tagged shipping information. The shipping information is provided by the well-known LLLOYD's lists dataset. We have introduced the case study and dataset characteristics used in the research project and presented our data model in Unified Modeling Language (UML). The model focuses on spatio-temporal events where characteristics are categorized as thematic, spatial and temporal attributes. The paper follows up with discussion on the selected approach and results, and finally ends with presenting the future outlook.

Keywords: Spatio-temporal, data modeling, Lloyd's lists

1 Introduction

Spatio-temporal data models deal with capturing information characterized by both spatial and temporal semantics. Research in this area started decades ago when collection and management of data, related to both spatial and temporal changes was recognized as an essential task. Before that, earlier works in this area began separately in spatial [1] and temporal [2] data models. These efforts later became the basis for generation of data models which can handle spatio-temporal context as a whole.

Nowadays, several data models have been presented regarding research and practice on spatio-temporal modelling. Each model has its own approach for dealing with spatio-temporal data and facing the challenges concerned with the integration of spatial and temporal concepts. Therefore, each model would be useful to be applied for a specific application (depending on the requirements of application as well as data characteristics) and not for another.

In this research, we aim to study the available spatio-temporal data models and select the best candidate for designing a spatio-temporal data model for geo-tagged shipping information. In the next section we provide the basic definitions and provide a brief review of the literature regarding spatio-temporal data models. In section 3, our case study and data is presented and based on discussion and interpretations, the relevant approach for modelling this data is selected. This section also provides our initial results of applying the approach, and modelling our data. Finally, in section 5, the paper ends with providing the discussion, conclusion and our plan for future work.

2 Spatio-temporal data modeling

Spatio-temporal data models define object data types, relationships, rules and constraints that maintain the integrity of a database [8]. A well-defined data model must anticipate spatio-temporal queries, geo-analytical and geo-statistical methods to be performed in a Geographical Information System (GIS). Their purpose is to deal with real world applications where spatial changes occur over time. The following four categories define the main requirements that need to be addressed by a spatio-temporal data model [8]:

- **Spatial semantics:** this category contains several criterions dealing with pure spatial aspects such as *structure of space*, *orientation/direction*, *measurement* and *topology*.
- **Temporal semantics:** this category deals with the nature of time and the basic features which are used to describe it such as *granularity*, *time density*, *time order*, *transaction*, etc. The issue of whether time should be modelled as continuous elements or as discrete elements are covered with these criterions.
- **Spatio-temporal semantics:** this category contains the most important and challenging criterions that do not exists in single spatial or temporal data models. *Data types*, *primitive notions*, *type of change*, *evolution in time and space*, *space-time topology*, *object identities*, and *dimensionality* are the seven factors/criterions which for instance, makes the data model able of capturing changes in shape and size of the object features and/or whether

* Corresponding author: Amin Mobasheri, E-mail: amin.mobasheri@geog.uni-heidelberg.de

a model supports spatio-temporal real world objects that change continuously or just objects that are subject to discrete changes.

- **Query capabilities:** this category is devoted to classification of existing spatio-temporal data models based on their query capabilities. Such queries may contain:
    - Queries about locations, spatial properties, and spatial relationships
    - Queries about time, temporal properties, and temporal relationships, and
    - Queries about spatio-temporal behaviours and relationships

However, the above queries form a minimum functionality that a spatio-temporal system should provide [3].

Several models exist for spatio-temporal data modelling. These models range from simple approaches such as The snapshot model [10], Simple Time-stamping [5], and Space-Time Composite (STC) [11] to more sophisticated models such as Entity-Relationship (STER) model [12], and Spatio-temporal Object-Oriented data models [13, 14].

The Snapshot model is one of the simplest spatio-temporal data models where temporal information has been incorporated into the model by time-stamping layers [10]. The model considers every layer as collection of temporally homogeneous units of one theme and shows the states of a geographic distribution at different times without considering temporal relations among layers, explicitly. This is the simplest way to capture spatio-temporal information and therefore has limitation in means of being capable to support complex queries.

Another approach is to tag every object with a pair of timestamps. One tag could be for the time of creation and the other for the time of cessation. This approach is also known as data models based on simple time-stamping. Previous implementation of such an approach shows that time slices can easily be retrieved by simple queries [5]. The authors argue that storing layers of geo-information for different time periods is impractical. Therefore, they develop a system that keeps a graphic file of current parcels for day to day use, but archive historical spatial data into a separate file and keep the references alive.

In addition to these approaches, there are other data models that follow the conceptual database models. The Spatio-Temporal Entity-Relationship (STER) model [6] is one of the earliest and best known models among them. The STER model is able to deal with complex geo-entity sets and interrelations of spatial and temporal semantics allows description of attributes and relationships among entity sets. The model is universal in terms of reusability because of its flexible notation [7, 8].

As another approach, spatio-temporal Object-Oriented data models incorporate the features of object oriented technology such as classes, instances, abstract data types, encapsulation, aggregation, inheritance, polymorphism, etc. In a research study, Wachowicz and Healy [9] present an object-oriented spatio-temporal model of real-world phenomena and events. The phenomena are represented as complex versioned objects with geometric, topological and thematic properties. In this approach, for every version of the object which establishes a hierarchical structure for the past, present and future of the object, a new instance of the object with a different identifier is created. In addition, "events are manifestations of actions, which invoke update procedures on one or more objects" [8]. In this approach, time is represented as an independent, linear dimension. The time reference is absolute and the time order is linear. Last but not least, space is conceptualised in three linear dimensions [9].

In practice, there is no "complete" model of any kind for any application domain. The decision for selecting the relevant approach of modelling should be made based on the requirements for the specific application as well as the characteristics of information aimed to be modelled. Therefore, in the next section we present our case study and thorough discussion select the most relevant approach to be applied for spatio-temporal modelling of geo-tagged shipping information in our research.

3 Case study: The LLOYD's Lists

Since the late seventeenth century, the shipping newspaper Lloyd's List and its direct predecessors contain weekly and later daily information on global shipping. The core of the Lists' mostly tabular contents is formed by the categories *"Shipping Intelligence"*, *"Speakings"*, *"Foreign Mail"*, *"Casualities"*, and *"War"*. Specifically, the first two categories are essential in our research. The "*Shipping Intelligence*" consists of exhaustive lists of the arrivals, departures and other nautical activities of civilian ships in practically all important ports of the world. The *"Speakings"* list sightings of ships at the high seas and give both the sighted and the reporting ship with name and geographical coordinates.

The *"Speakings"* and *"Intelligence"* hold much information that with a primarily quantitative approach, will allow us to analyze, for instance, the shifting patterns of shipping routes; the time it took to get information, goods, and humans from one harbor to another depending on the year and season; the "black spots" and interruptions of service due to natural disasters or wars; the shifts in trade intensity between specific regions; the constantly changing patterns of transcontinental/international trade and migration; or, in short, the transformation of a variety of "global spaces" during times of rapid globalization [15].

As an initial step, spatio-temporal modeling of Lloyd's data is essential in order to capture the information in a coherent and flexible manner which would later allow spatio-temporal analysis and querying of this information. The *"Speakings"* information contains latitudes and longitudes that could be best captured by a point feature as well as two timestamps. The first timestamp is for the date that the actual *speaking* has occurred and the other timestamp contains the actual date when this information has been reported.

In order to deal with modeling of Lloyd's data, some approaches cannot be used due to their specific drawbacks. For example, the Snapshot model is not appropriate for means of describing changes in space thorough time. Each snapshot is relevant to a specific time, but in order to understand how T_i differs from T_j, two snapshots should be compared exhaustively. As an example of disadvantages of the Space-

Time Composite (STC) model, the fact that it is very difficult to define rules of internal logic and/or integrity constraints is a big problem. This is because the model does not provide understanding of the constraints upon the temporal structure [8]. Employing the STER model on the other hand has several benefits but the main problem according to a research study [8], which makes this model un-suitable for our research too, is the fact that it lacks the ability to capture the actual motion of the process of change and does not indicate if a spatial object is dynamic or static.

Among several approaches of spatio-temporal data modeling discussed in the previous section, we select the object oriented modeling because of its four main advantages in spatio-temporal modeling [8, 13]:

- A single object represents the whole history of an entity
- Efficient temporal data handling
- Uniform treatment of spatial and temporal data handling
- Simple queries due to its capability of dealing with each single object of an entity

As an initial attempt towards digitization, a large amount of published information for specific time periods which was of interest for historians; specifically years 1851 and 1871 were read and transferred into Excel spreadsheets in a formal tabular format. In the next step, in order to design a geo-database for the Lloyd's data, a geo-data modeling task using the object oriented modeling was performed. As an extra data source, we will also use the digitized maps of shipping routes downloaded from David Rumsey's map collection [17] and the CLIWOC database [16] which provide weather data (e.g. wind speed and direction, air temperature, etc.). The aim would be to integrate these datasets with the Lloyd's data in order to improve the normal trajectories (digitized from the maps) to a higher abstraction level, leading to semantic trajectories. From all the necessary information that needed to be recorded, a total amount of 9 tables were designed, each of which carry special attributes.

Figure 1 illustrates the data model for each table in a class diagram fashion using Unified Modeling Language (UML). Note that the first four tables are normal tables containing several necessary information (e.g. Ship name, Captain name, Event date, Journal Date, etc.), yet the two last tables are considered as feature types since they have spatial components, thus can be treated as point features. As it can be seen two main tables are *tbl_Speakings* and *tbl_Intelligence* which contain the shipping information. Furthermore, *tbl_Speakings* is the most special table in this model which contains coordinate values in forms of latitude and longitudes where the actual speaking has happened (somewhere in the sea) as well as information about the plan of the other ship (the sighted ship) such as its port of origin and destination. *tbl_weather* captures weather information collected from CLIWOC database [16] for the relevant temporal periods of LLOYD's shipping data. Last but not the least, *tbl_Routes* captures the data model behind shipping routes that are digitized based on famous maps provided by David Rumsey's map collection [17] (note that additional data are stored in *tbl_Route_Metadata*).

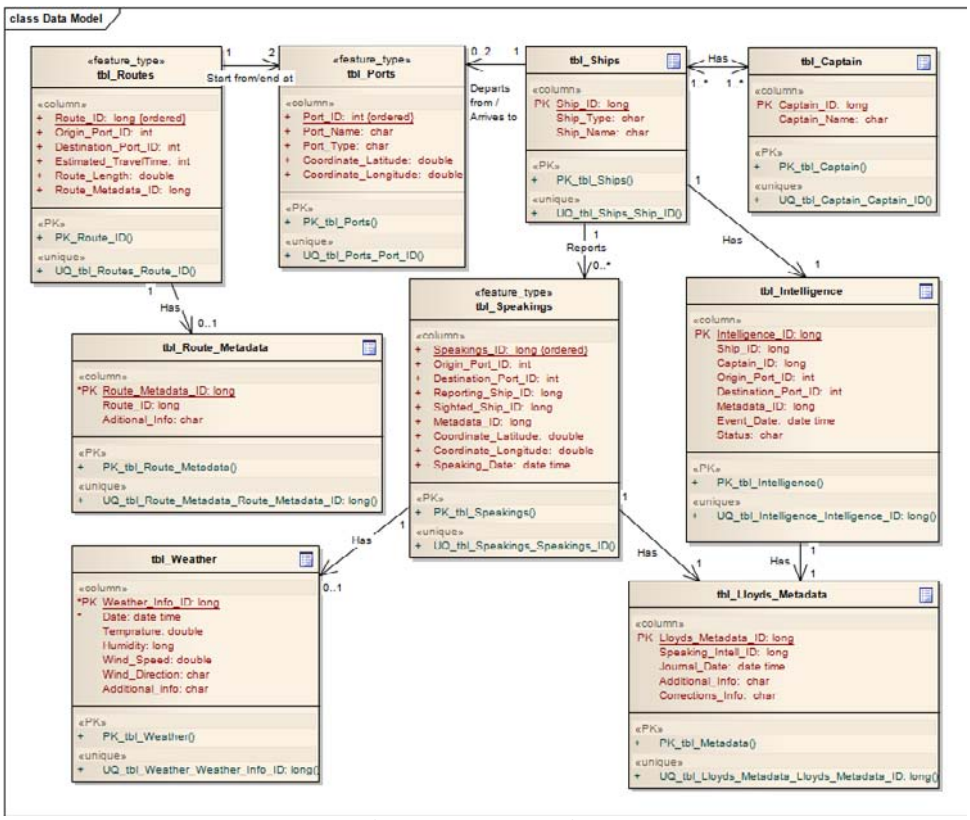


Figure1: Main part of the data model in Unified Modeling Language

The model focuses on spatio-temporal events where characteristics are categorized as thematic, spatial and temporal attributes. In this approach, for every version of the object which establishes a hierarchical structure for the past, present and future of the object, a new instance of the object with a different identifier is created. In addition, "events are manifestations of actions, which invoke update procedures on one or more objects". In our model, speakings of ships are the special events that occur in the sea and need to be recorded in an efficient and effective manner. In this approach, time is represented as an independent, linear dimension. The time reference is absolute and the time order is linear. Last but not least, space is conceptualized in two linear dimensions.

4 Discussion and Conclusion

In this paper we presented a brief overview of available approaches for spatio-temporal data modeling and elaborated on some of the advantages and disadvantages of them. In the next step, we introduced the data of our case study: Lloyd's lists, and through a discussion selected the object oriented modeling as the best candidate for spatio-temporal modeling of geo-tagged shipping information. The paper also presents our initial attempt to design our model using Unified Modeling Language (UML). We concluded that our model focuses on spatio-temporal events and for every version of the object which establishes a hierarchical structure for the past, present and future of the object, a new instance of the object with a different identifier is created. *Speakings* of ships are seen as specific events happening in the sea with additional semantic information and temporal attributes that need to modelled effectively and efficiently in order to be ready for spatio-temporal analysis and reasoning tasks.

Furthermore, using simple sample points would not be the only case in this research, we will also use a higher level of abstraction in order to introduce semantic trajectories of data. The reason is that more application-oriented ways of analyzing segments of movement suitable for specific purposes of the application domain is needed. Therefore, in the future, the model provided here would be extended in order to be capable of capturing trajectories (e.g. lines) in addition to point features. Another aim for future work is design and populate a geo-database based on the data model and to review and apply relevant geo-statistical methods for means of analyzing and discovering spatio-temporal patterns of shipping information and detecting hot-spots of shipping *speakings* considering different temporal timestamps.

5 Acknowledgments

This research is made possible by a grant from the Cluster of Excellence "Asia and Europe in a Global Context" as well as GIScience research group of University of Heidelberg. Also, the authors would like to thank the hard work of students in the cluster as well as the department of history (HGIS Club) for extracting the data samples of the Lloyd's lists. The authors would also like to thank Alexander Zipf, Kilian Schultes and Roland Wenzlhuemer from Heidelberg University for their comments and support.

References

[1] R.H. Guting, "An Introduction to Spatial Database Systems", VLDB Journal 4 (1994), 357-399.

[2] A.U. Tansel, J. Clifford, S. Gadia, S. Jajodia, A. Segev and R. Snodgrass, "Temporal Databases: Theory, Design and Implementation", Benjamin/Cummings Publishing Company, 1993.

[3] M. Koubarakis, T. Sellis et al. (eds.), Spatio-Temporal Databases: The Chorochronos Approach, 2003. Springer-Verlag LNCS 2520.

[4] G. Langran, "A Framework for Temporal Geographic Information Systems", Cartographica, 25 (3), 1988.

[5] G. J. Hunter and I. P. Williamson, "The Development of a Historical Digital Cadastral Database", Int. Journal of Geographic Information Systems, 4(2), 1990.

[6] N. Tryfona and C. S. Jensen, "Conceptual Data Modeling for Spatiotemporal Applications", GeoInformatica, Vol. 3: 245-268, 1999.

[7] N. Tryfona and C. S. Jensen, "Using Abstractions for Spatio-Temporal Conceptual Modeling", Proceedings of the 2000 ACM Symposium on Applied Computing, Como, Italy, 2000.

[8] P. Nikos, B. Theodoulidis, I. Kopanakis, and Y. Theodoridis. "Literature review of spatio-temporal database models." The Knowledge Engineering Review 19, no. 03 (2004): 235-274.

[9] M. Wachowicz and R. G. Healey, "Towards Temporality in GIS", In Worboys M. F. editor, Innovations in GIS, Taylor & Francis, 1994.

[10] G. Langran, "A Framework for Temporal Geographic Information Systems", Cartographica, 25 (3), 1988.

[11] G. Langran, "Time in Geographical Information Systems", ed. Taylor & Francis, London, 1992.

[12] I. Theodoulidis and P. Loucopoulos, "The Time Dimension in Conceptual Modeling", Information Systems, vol. 16, no. 3, pp. 273-300, 1991.

[13] M. F. Worboys, H. M. Hearhshow, D. J. Maguire, "Object-Oriented Modeling for Spatial Databases", Int. Journal of GIS ol. 4, No. 4, 1990.

[14] M. Wachowicz and R. G. Healey, "Towards Temporality in GIS", In Worboys M. F. editor, Innovations in GIS, Taylor & Francis, 1994.

[15] A. Mobasheri, M. Bakillah, K. Schultes, A. Zipf. Towards Integrated Web Processing Services for Spatio-temporal Analysis of Geo-data, The Case of Lloyd's Lists. Scientific Computing and Cultural Heritage Conference (SCCH), Heidelberg, Germany, 2013.

[16] Wheeler, D., Garcia-Herrera, R., Koek, F.B., Wilkinson, C., Können, G.P., Prieto, M.R., Jones, P.D. and R. Casale. 2006: CLIWOC, Climatological database for the world's oceans: 1750 to 1850; Results of a research project EVK1-CT-2000-00090. European Commission, Brussels, ISBN 92-894-8279-6.

[17] David Rumsey Map Collection, available at http://www.davidrumsey.com/view, accessed in May 2014.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Towards initiating OpenLandMap founded on citizens' science: The current status of land use features of OpenStreetMap in Europe

Jamal Jokar Arsanjani
GIScience research group
Heidelberg University
Berliner str. 48, 69120
Heidelberg, Germany
jokar.arsanjani@geog.uni-heidelberg.de

Eric Vaz
Department of Geography
Ryerson University
evaz@geography.ryerson.ca

Mohamed Bakillah
GIScience research group
Heidelberg University
Berliner str. 48, 69120
Heidelberg, Germany
mohamed.bakillah@geog.uni-heidelberg.de

Peter Mooney
Department of Computer Science,
National University of Ireland, Maynooth
peter.mooney@nuim.ie

Abstract

Land use inventories are important information sources for scholarly research, policy-makers, practitioners, and developers. A considerable amount of effort and monetary resources have been used to generate global/regional/local land use datasets. While remote sensing images and techniques as well as field surveying have been the main sources of determining land use features, in-field measurements of ground truth data collection for attributing those features has been always a challenging step in terms of time, money, as well as information reliability. In recent years, Web 2.0 technologies and GPS-enabled devices have advanced citizen science (CS) projects and made them user-friendly for volunteered citizens to collect and share their knowledge about geographical objects to these projects. Surprisingly, one of the leading CS projects i.e., OpenStreetMap (OSM) collects and provides land use features. The collaboratively collected land use features from multiple citizens could greatly support the challenging component of land use mapping which is in-field data collection. Hence, the main objective of this study is to calculate the completeness of land use features to OSM across Europe. The empirical findings reveal that the completeness index varies widely ranging from almost 2% for Iceland to 96% for Bosnia and Herzegovina. More precisely, more than 50% of land use features of eight European countries are mapped. This shows that CS can play a role in land use mapping as an alternative data source, which can partially contribute to the existing inventories for updating purposes.

Introduction

Land use/cover maps are essential for environmentalists and land managers for urban and regional planning purposes. These maps identify which features exist on the ground and for which purpose each land parcel is used [26,32]. The process of mapping land related features is called land use/cover mapping e.g., [23,34], which result in land use/cover inventories. Traditionally land surveying and recently remote sensing data and algorithms have been used to map land use/cover patterns e.g., [22,28,30]. Undoubtedly, remote sensing has played a vital role in monitoring and mapping land features. Nevertheless, in-field information is often required to assess the outcomes of remote sensing techniques [3,5]. Additionally, they are used to enrich the land use patterns regarding its attributes and semantic information [13].

Recently, the rise of web 2.0 technologies and CS-based projects has resulted in tremendous amount of geolocated information from citizens [9,16]. As a successful leading CS projects, OSM can be named, which has been increasing receiving new users and contributions. Published investigations on applicability of OSM datasets have shown that OSM provides us a wide variety of datasets for different application including and not limited to routing, Points of Interest (POIs) search, transport mapping, building inventories, etc. OSM also collects the information on land features and shares them with public. So far, little attention to the collected OSM features on land use information has been drawn [4,8], although OSM can provide an alternative source for mapping land use features contributed by citizens. What is remarkable about harnessing OSM for land use mapping is the fact that once OSM users log into OSM, fine resolution image libraries generated from multiple remote sensing imageries are shared in the mapping/editing interface so that the users simply delineate the geometrical tessellation of land use features and additionally insert their personal knowledge of that specific land parcel to it. It is of great importance to note that in this process, the OSM users benefit from user-friendly editing softwares, which display fine-resolution images (even up to 20 cm spatial resolution) in the background, for delineating land parcels and add attributes and metadata about each land parcel to it [21]. In other words, thanks to the fine-resolution images/air-photos as well as users' knowledge of the mapped areas, the process of land use mapping is handled differently so that the in-field information are actively given by the users instead of going to the field for collecting them [20].

A remarkable amount of efforts and money have been inserted into generating global land-use maps, for instance, Global Land Cover (GLC)-2000 [11], Moderate-resolution Imaging Spectroradiometer (MODIS; [10]), and GlobCover [1], among others. At a European level scale, the CORINE 2000 [2] and Global Monitoring for Environment and Security Urban Atlas (GMESUA; [3]) have been prepared. The accuracy of these inventories however, is often questioned by the researchers and further projects on evaluating their accuracies are called [19,25,27,29,33]. To sum up, the process of generating land use inventories actively demands for large amount of budget, while this process in a passive manner diminishes the monetary costs significantly and might result in better results. Furthermore, they need to be updated on a regular basis and therefore, repeating the efforts. As such, the main aim is to evaluate the degree of completeness for OSM land use features in order to see how well OSM can play a role in land use science. Empirical findings reported by [15,20] have addressed the potentials of exploiting OSM for land use mapping. Hence, the main objective of this study is to measure how complete OSM land use features in a European scale are in order to start exploiting them. To be

more precise, this research seeks to find out how complete land-use features per each European state are contributed to OSM.

Materials and data processing

3.1 OpenStreetMap dataset

The OSM datasets utilized in this study is the OSM snapshot for February 20, 2014. To retrieve relevant land-use features, A country-wide coverage of forty European countries is sampled in this study. The reason for considering a pan-European wide of datasets is the fact the patterns of contributions are intrinsically heterogeneous as proven by [17,21]. This is also evident through a query to osmatrix.uni-hd.de. Figure 1 displays the extent of this study.

Figure 1: the selected study areas

Methods

Among the purposed criteria by different ISO standards in particular 19157:2013 for assessing the accuracy of geodata internally, completeness plays a vital role as it measures how complete the dataset is [7,14]. Completeness is the major concern for using OSM datasets [18,24] as it is an indicator of how much of the whole has been mapped by volunteers. In contrast to polyline and point features in OSM, the completeness for land-use features is the proportion of mapped areas relate to its overall extent. The completeness index for each country is calculated by calculating the mapped areas by the whole area of extent. This represents a simple indicator to find out how complete a country is mapped i.e., how far we are from having full data coverage.

Results and discussion

Table 1 represents total mapped area and completeness indices for each country. As shown in Table 1, the calculated completeness index values are diverse. While only 1.6% of land use features in Iceland are mapped, 96% of Bosnia and Herzegovina are mapped, which is quite surprising that no study has been already dedicated to further accuracy assessment of the contributed features.

Table 1: the calculated completeness values for each country

Country	Total Area (km2)	Mapped Area (km2)	Completeness (%)	Class
Bosnia & H.	51,209	49,495	96.6	A
Slovakia	49,035	43,698	89.1	A
Netherlands	37,354	30,818	82.5	A
Belgium	30,528	19,221	63.0	A
Romania	238,391	138,737	58.2	A
Luxemburg	2,586	1,426	55.2	A
France	548,500	296,833	54.1	A
Germany	357,114	190,851	53.4	A
Liechtenstein	160	65	41.2	B
Macedonia	25,713	9,432	36.7	B
Czech R.	78,867	28,728	36.4	B
Croatia	56,594	17,591	31.1	B
Andorra	468	144	30.9	B
Poland	312,685	88,489	28.3	B
Austria	83,945	22,764	27.1	B
Denmark	43,094	11,610	26.9	B
Switzerland	41,277	10,803	26.2	B
Cyprus	9,251	2,422	26.2	B
Slovenia	20,273	5,240	25.8	B
Finland	338,419	86,569	25.6	B
Montenegro	13,812	2,916	21.1	B
Spain	505,992	106,131	21.0	B
Greece	131,957	27,181	20.6	B
Great Britain	242,900	46,366	19.1	B
Lithuania	65,300	12,108	18.5	B
Kosovo	10,908	2,004	18.4	B
Norway	386,224	61,706	16.0	B
Moldova	33,846	5,410	16.0	B
Malta	316	48	15.4	B
Hungary	93,028	14,198	15.3	B
Serbia	88,361	11,481	13.0	B
Bulgaria	110,879	14,362	12.9	B
Sweden	441,370	56,657	12.8	B
Italy	301,336	38,024	12.6	B
Ukraine	603,500	68,735	11.4	B
Belarus	207,600	22,968	11.1	B
Ireland	70,273	4,965	7.1	B
Portugal	92,090	3,919	4.3	B
Albania	28,748	897	3.1	B
Iceland	103,000	1,687	1.6	B

polygon features labelled with “Land-use” and “Natural” tags are filtered. While the features with “Natural” tag describe a wide variety of physical features, features with “Land-use” tag identify the land use features. These features are then merged together to create a uniform dataset.

3.2 Study area

The completeness indices are then arbitrarily categorized into two classes ranging between zero to hundred percent with 50 percent interval. To be more precise, while class “A” represents countries that completeness index exceeds 50 percent, class “B” identifies countries that less than half of them are mapped. According to this categorization, 8 countries place within the class “A” and 32 countries are classified as “B”. Belgium, Bosnia & Herzegovina, Germany, France, Luxemburg, the Netherlands, Romania, and Slovakia are those which are well-mapped. Spatial distribution of the mapped features within Europe is displayed in Figure 2. Green cells represent the contributed features regardless their attributes. It should be mentioned that the European countries have different populations and population densities, and physical characteristics and the completeness values should not be used for refereeing the topology of citizen participations in collaborative mapping practices [17]. For instance, Iceland with an area of 103,000 km^2 and nearly 300 thousand inhabitants is the least mapped country. This is not comparable with the Netherlands, holding an area of 41,500 km^2 and nearly 17 million inhabitants, corresponding to the best mapped country (82%). This inequality of public participation should be further investigated.

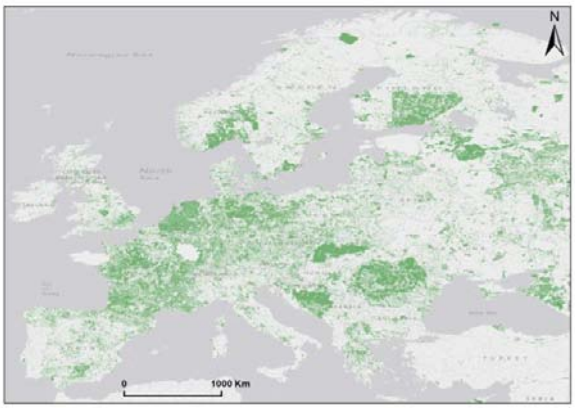


Figure 2: spatial distribution of contributed land use features in Europe

Conclusion

The contemporary emergence of citizen science projects, namely OSM, has drawn the attention of large number of citizens to share their information, as well as records of their GPS-enabled devices, with the public. This collaboratively collected information have been implemented in several applications such as navigation, context-aware routing, indoor mapping, and tourism recommendations. Exceptionally, OSM collects the land use features from contributors and therefore its potential for land use science has to be assessed.

This study aimed at assessing the completeness of land use features across European countries to find out how completely these features have been mapped. The calculated indices reveal that the degree of completeness is heterogeneous and ranges between 1 to 96 percent. More than half of 8 countries as listed in Table 1 are mapped in terms of land use features by OSM mappers. Apart from barely mapped countries, this means that volunteered mappers express their interest in mapping landscape related information as well and this opens avenues for further research towards harnessing CS for land use science. Future research directions should be conducted towards accuracy assessment of the land use attributes versus ground truth or proprietary datasets, e.g., the pan-European urban atlas and CORINE datasets.

As a final conclusion, the contributed OSM land use information suggest a promising alternative data source for land use mapping independent from applying computational image processing techniques. Whereas the degree of completeness in OSM increases over time, further contributions from volunteers should be expected within a short period of time. Further to this, the findings attempt to draw the attentions of volunteers to map the landscape-related objects as well so that citizen science could greatly contribute to collecting up-to-date information of our land resources. The following recommendations are suggested to environmentalists and land-use scientists that contributed features enable us to either consider the OSM features as an alternative data source or take advantage of the partially mapped areas for updating the existing and outdated inventories as outlined by [12]. It should be mentioned that applying data mining and data fusion techniques with other available features in OSM help to complete the incomplete areas.

References

[1] Bontemps, S., Defourny, P., Van Bogaert, E., Arino, O., Kalogirou, V., Ramos, P., Jose, J., (2011). GLOBCOVER 2009 Products description and validation report. Université catholique de Louvain (UCL) & European Space Agency (ESA), Vers. 2.2, pp 53.

[2] Buettner, G., Feranec, J., and Jaffrain, G., (2002). Corine land-cover update 2000. EEA. Technical Report, vol. 89, 17 December 2002. Copenhagen.

[3] Cihlar, J. and Jansen, L.J.M., (2001). From land-cover to land-use: a methodology for efficient land-use mapping over large areas. The Professional Geographer, 53 (2), 275–289.

[4] Comber, A., Brunsdon, C., See, L., Fritz, S., and McCallum, I. (2013). Comparing Expert and Non-expert Conceptualisations of the Land: An Analysis of Crowdsourced Land Cover Data. In T. Tenbrink, J. Stell, A. Galton, & Z. Wood (Eds.), Spatial Information Theory SE - 14 (Vol. 8116, pp. 243–260). Springer International Publishing.

[5] De Leeuw, J., Said, M., Ortegah, L., Nagda, S., Georgiadou, Y., & DeBlois, M. (2011). An assessment of the accuracy of volunteered road map production in Western Kenya. Remote Sensing, 3(2), 247-256.

[6] Devillers, R., and Jeansoulin, R., (2006). Fundamentals of Spatial Data Quality, ISTE, London.

[7] Devillers, R., Bédard, Y., Jeansoulin, R., and Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data, International Journal of Geographical Information Sciences, 21(3), 261–282.

[8] Estima, J., & Painho, M. (2013). Exploratory analysis of OpenStreetMap for land use classification. In Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (pp. 39-46).

[9] Foody, G.M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., and Boyd, D. S., (2013). Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. Transactions in GIS, 17(6), pp. 847–860.

[10] Friedl, M.A., McIver, D.K., Hodges, J.C., Zhang, X.Y., Muchoney, D., Strahler, A.H., and Schaaf, C., (2002). Global land cover mapping from MODIS: algorithms and early results. Remote Sensing of Environment, 83(1), 287-302.

[11] Fritz, S., Bartholomé, E., Belward, A., Hartley, A., Stibig, H. J., Eva, H., Mayaux, P., ... and Defourny, P. (2003). Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version) (p. 41). Luxembourg: Office for Official Publications of the European Communities.

[12] Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., Obersteiner, M., (2009). Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. Remote Sensing, 1, 345-354.

[13] Fritz, S., Mccallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., … Obersteiner, M. (2012). Environmental Modelling & Software Geo-Wiki: An online platform for improving global land cover. Environmental Modelling and Software, 31, 110–123.

[14] Gervais, M., Bédard, Y., Levesque, M., Bernier, E., and Devillers, R., (2009). Data Quality Issues and Geographic Knowledge Discovery, 99–116.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

[15] Hagenauer, J., and Helbich, M., (2012). Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. International Journal of Geographic Information Science, 26 (6), 963–982.

[16] Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design, 37(4), 682–703.

[17] Haklay, M., (2013), Citizen Science and Volunteered Geographic Information – overview and typology of participation in Sui, D.Z., Elwood, S. and M.F. Goodchild (eds.), Crowdsourcing Geographic Knowledge. Berlin: Springer. pp. 105-122

[18] Hecht, R., Kunze, C., and Hahmann, S., (2013). Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time. ISPRS International Journal of Geo-Information, 2(4), 1066–1091

[19] Herold, M., Mayaux, P., Woodcock, C. E., Baccini, A., & Schmullius, C., (2008). Some challenges in global land-cover mapping: An assessment of agreement and accuracy in existing 1 km datasets, Remote Sensing of Environment 112(5), 2538–2556.

[20] Jokar Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., and Zipf, A., (2013a). Toward mapping land-use patterns from volunteered geographic information, International Journal of GIS, 27(12), 2264-2278.

[21] Jokar Arsanjani, J., Helbich, M., Loos, L., Bakillah, M., (2014:in-press). The emergence and evolution of OpenStreetMap: A cellular automata approach, International Journal of Digital Earth.

[22] Kandrika, S. and Roy, P.S., (2008). Land-use land-cover classification of Orissa using multi-temporal IRS-P6 AWIFS data: a decision tree approach. International Journal of Applied Earth Observation and Geoinformation, 10 (2), 186–193.

[23] Kasetkasem, T., Arora, M. K., & Varshney, P. K. (2005). Super-resolution land cover mapping using a Markov random field based approach. Remote Sensing of Environment, 96(3–4), 302–314.

[24] Koukoletsos, T., Haklay, M., and Ellul, C., (2012). Assessing data completeness of VGI through an automated matching procedure for linear data. Transactions in GIS, 16 (4), 477–498.

[25] Mayaux, P., Eva, H., Gallego, J., Strahler, A. H., Herold, M., Member, S., Agrawal, S., (2006). Validation of the Global Land-cover 2000 Map, IEEE Transactions on Geoscience and Remote Sensing, 44(7), 1728–1739.

[26] Paneque-Gálvez, J., Mas, J.-F., Moré, G., Cristóbal, J., Orta-Martínez, M., Luz, A. C., Reyes-García, V., (2013). Enhanced land use/cover classification of heterogeneous tropical landscapes using support vector machines and textural homogeneity. International Journal of Applied Earth Observation and Geoinformation, 23, 372–383.

[27] Pontius, R.G. Jr., and Petrova, S.H., (2010). Assessing a predictive model of land change using uncertain data. Environmental Modelling & Software, 25 (3), 299–309.

[28] Qi, Z., Yeh, A. G. O., Li, X., and Lin, Z., (2012). A novel algorithm for land use and land cover classification using RADARSAT-2 polarimetric SAR data. Remote Sensing of Environment, 118, 21-39.

[29] Robinson, D.T. and Brown, D.G., (2009). Evaluating the effects of land-use development policies on ex-urban forest cover: an integrated agent-based GIS approach. International Journal of Geographical Information Science, 23 (9), 1211–1232.

[30] Saadat, H., Adamowski, J., Bonnell, R., Sharifi, F., Namdar, M., and Ale-Ebrahim, S. (2011). Land use and land cover classification over a large area in Iran based on single date analysis of satellite imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 66(5), 608–619.

[31] Seifert, F., (2009). Improving Urban Monitoring toward a European Urban Atlas. In Global Mapping of Human Settlement. CRC Press.

[32] Sexton, J. O., Urban, D. L., Donohue, M. J., and Song, C. (2013). Long-term land cover dynamics by multi-temporal classification across the Landsat-5 record. Remote Sensing of Environment, 128, 246–258.

[33] Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., Mayaux, P., (2006). Global Land-cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land-cover Maps. Luxemburg : Office for Official Publications of the European Communities, (25).

[34] Thenkabail, P.S., Schull, M., and Turral, H., (2005). Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data. Remote Sensing of Environment, 95(3), 317–341.

Session:
Observation Integration

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Augmented Reality and GIS: On the Possibilities and Limits of Markerless AR

Falko Schmid and Daniel Langerenken
Cognitive Systems
University of Bremen
Germany
schmid@informatik.uni-bremen.de, daniel.langerenken@gmail.com

Abstract

The application of Augmented Reality (AR) in the geo-spatial domain offers huge potentials: AR can visualize invisible properties of spatial entities, can display historic data for them, or can help in finding places. Whatever the application is, AR in the geo-spatial domain will often be purely sensor based, thus without the help of visual or sensory markers. In this paper we analyse the achievable accuracy of AR projections under everyday conditions with consumer hardware. We can show that AR can be applied in applications in smaller geographic scale, but is not sufficient if it comes to the preciseness required when inspecting infrastructural data of small scale.

Keywords: AR, GIS

1 Introduction

Enriching the direct perceivable environment with complementary information bears great potential in the geo-spatial domain. We can make the invisible visible, we can browse through history and future of a place, we can learn about legal issues, we can assist during navigation, advertise properties, etc. With augmented reality (AR) we can visualize the road to take, underground pipe and cable installations, the type of soil below us, its quality, and contamination with toxic substances. We can learn about archaeological discoveries of filled up digging sites, see the places that have been flooded or will be at a certain water level, or how buildings will look like when they are built.

The possibilities are endless and with the broad availability of sensor-packed devices like smartphones and the advent of data glasses in the end-user market, augmented reality (AR) will be the tool of choice for many of these applications. AR applications can help to make informed decisions, reduce costs, entertain, and assist during spatial tasks. However, this is only possible if the applications can support the required level of accuracy. I.e., accurate projections are required to ensure that projected data corresponds with the entities of the camera image. The level of required accuracy depends on the domain: some applications will be usable even if the results are displaced by 10 meters, others will require a high degree of precision.

Projecting data at the correct camera image technically requires accurate positioning, clear sensory data, and ideally some visual or sensory makers for precise alignment of data in the environment. State-of-the-art techniques ensure accuracies down to millimetre precision, this level of accuracy will be out of reach for the majority of geo-spatial applications for the next years. High precision can be achieved in constrained domains and controlled settings where the system knows about clear markers, visual properties of environments and entities, or has access to precise sensors. Although precision and availability of technology constantly increase positioning and 3D orientation sensing will have limited accuracy in everyday settings and away from lab conditions.

GPS-based positioning with non-survey grade devices is known to be inaccurate, Wi-Fi is and will not be available everywhere in the world, and the environment is constantly changing due to evolution, seasonal features, or events. Landmarks, buildings, signs, trees, and parks appear and disappear. Thus, the available data, which is the potential source for sensory or visual registration methods can differ significantly from reality: the building an algorithm is looking for can be replaced, the street can be covered with snow, and the tree is currently without leafs.

AR literature and its evaluations suggest that that markerless, pure sensor-based AR is not sufficient for applications requiring high precision projections. However, this is certainly true for applications requiring a high degree of precision (e.g., surgical applications) - for other classes of applications the limitations might be acceptable. In this paper we analyse the limits of pure sensor-based, markerless AR under everyday conditions and identify classes of applications suitable for the achievable accuracy.

2 Related Work

During the last years the application context of AR-based applications strongly moved to the direction of the broad mass of users. Due to technically very powerful and affordable smartphones and the possibilities of developing your own mobile applications, more and more applications are published that mix real and virtual environment. Liarokapis et al. justify this by the rise of GIS. Therefore they developed a tangible user interface for visualizing geographical data received by shape files [1]. Another source of geodata is shown by Schmid et al in mapIT [2, 16]. They provided a

possibility to gather, annotate and send geodata to a GIS by using camera, sensor- and positioning data of smartphones.

Behringer linked sensor and positioning data with the image of a camera and height maps to register horizontal silhouettes in the viewport. This, however, requires good lighting conditions [3]. Stricker and Kettenbach describe an approach based on markerless, optical tracking. Depending of the current field of view of the camera, a collection of reference pictures is pre-sorted. From these images, the best reference image is calculated and then projected onto the camera image. Though, a known environment is needed to pre-sort a collection of reference pictures [4].

Azuma, Hoff et al. took care of the problem of inaccurate data and therefore developed a motion-stabilized outdoor AR-system. This system stabilizes the received sensor data and attempts to avoid delays by predicting. However, it is subject to some limitations due to the needed equipment. A fixed location is required to stabilize the received data. Changes in the location are not supported [5].

Yi Wu et al. studied the possibilities of outdoor AR in cities under consideration of the position, the orientation of the device and the current camera image. They linked sensor-based AR with natural marker-based AR. A database provided the necessary information for the current GPS position. [6].

For maintenance support Roberts et al. presented an AR-application which allowed to project gas, telephone, water and power lines located behind walls into the environment [7]. A similar approach is described by Behzadan et al. in projecting construction graphics into the real world [8]. They developed an AR-application, equipped with a HMD, a GPS receiver and a portable computer. The aim was to combine virtual reality with the construction, while the user is able to move freely in the environment.

Veas et al. investigated possibilities to extend the viewport in AR applications under different circumstances. Therefore they described the multiview-AR and variable-perspective-view. Thus, the user was able to see the field of view from different perspectives without the requirement to move. Moreover it is possible to swap between the first-person-view and a third-person-view to change the perspective variable [9].

Considering planar objects from a distance, thus causing the perspective projection to display objects in very small sizes which causes them to be very difficult to detect. This problem is known as “long flat view”, studied by King et al. [10]. One possible solution was to use a second camera, which is twice as high as the user. This doubles the field of view and therefore provides improved data for the depth. In addition to this problem King et al. studied also the problem of unreadable displays due to high solar radiation. This problem could be minimized by the use of dark, semi-transparent plastic on the screen or the use of umbrellas or hats. Also discussed was the issue of transparency of objects that are either not visible at certain color values during sunlight or they mask the reality completely.

In addition to the projection of objects there also exists the possibility to make objects disappear. This approach was described by Avery et al. [11]. In this case a mobile roboter was used to record hidden areas and transferring them directly to the user. Similar approaches to project hidden objects have been investigated by Webster et al. [12].

However, for most approaches it remains unclear which precision can be achieved under nowadays everyday conditions. Most approaches were tested under laboratory conditions, are marker-based, or hardware and software reality have changed drastically during the last years. In this paper we provide a glimpse on achievable accuracy under everyday conditions with standard AR projection techniques and consumer devices.

3 MapAR: An AR Tool for Geo-Data

In this paper we present MapAR, an AR tool for projecting invisible data or properties (e.g. collected by OpenStreetMap) in the camera image of everyday smartphones. With MapAR we are also evaluating the feasibility of markerless AR in context of geographic applications.

3.1 System Design

MapAR provides the possibility to project invisible data or properties in the camera image. Therefore it requires the coordinates of the data to be displayed. Figure 1 shows the projection of a parking lot in the main view of the application.

Figure 1: Arrow pointing to a parking lot. In MapAR.

3.2 Projection

Within MapAR we implemented following projection. To calculate screen coordinates, the position and orientation of the camera is required. Also the object to be projected must be available in Cartesian coordinates. Subsequently this data is used for a camera transformation to move the camera into the origin of the coordinate system. Thus, the coordinate system has to be rotated around the camera orientation $(\theta_x, \theta_y, \theta_z)$. As a result we obtain the point P (d_x, d_y, d_z) in camera coordinates [13, 14]. Figure 2 shows the corresponding matrix operation where (p_x, p_y, p_z) the current point to be projected is illustrated [16]. The first matrix causes the necessary rotation about the x-axis, the second matrix for rotation about the y-axis, and the third matrix of the rotation around the z-axis. Subsequently, the position of the camera from the point to be projected is subtracted to determine the position of the point in the camera system. Due to the perspective projection, we obtain a point in the camera coordinate system $(B(x, y, z))$.

The next step translates the obtained point B in screen coordinates. Therefore the viewport of the camera as well as the size of the screen ($width * height$) is required. The focal

$$\begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_z & -\sin\theta_z \\ 0 & \sin\theta_z & \cos\theta_z \end{bmatrix} \begin{bmatrix} \cos\theta_z & 0 & \sin\theta_z \\ 0 & 1 & 0 \\ -\sin\theta_z & 0 & \cos\theta_z \end{bmatrix} \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \left(\begin{bmatrix} p_x \\ p_y \\ p_z \end{bmatrix} - \begin{bmatrix} c_x \\ c_y \\ c_z \end{bmatrix} \right)$$

Figure 2: Calculation of screen point x

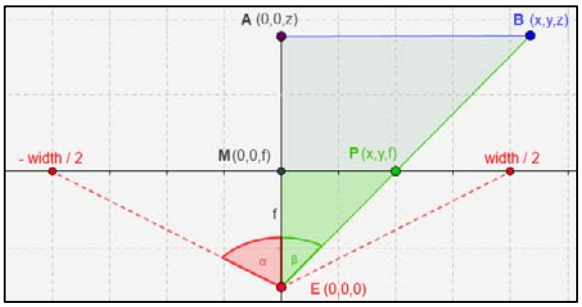
length, so the distance from the camera center to the projection area, can be calculated through trigonometric calculations. In the figure (2) the focal length is displayed by f. For the calculation of f, the horizontal view angle α and the width of the screen ($width$) is required:

$$\tan\alpha = \frac{\frac{(width)}{2}}{f} \tag{1}$$

Equation (1) can be resolved to f:

$$f = \frac{\frac{(width)}{2}}{\tan\alpha} \tag{2}$$

Figure 2: Calculation of screen point x



By using the side-splitter-theorem the corresponding screen position can be calculated. The side-splitter-theorem states that a line that is parallel to a side of a triangle and intersects the other two sides of the triangle, divides the area of the triangle proportional. Figure (2) shows the triangle ABE. This triangle is divided by $\overline{MP}$. In addition to that the line $\overline{MP}$ is parallel to the line $\overline{AB}$. The following applies:

$$\frac{E * P}{P * B} = \frac{E * M}{M * A} \tag{3}$$

By substituting the values of Figure (2) in Equation (3), we get:

$$\frac{x}{B(x)} = \frac{f}{B(z)} \tag{4}$$

By substituting f with the calculated focal length in Equation (4) we get:

$$\frac{x}{B(x)} = \frac{\frac{width}{2}}{\frac{\tan\alpha}{B(z)}} \tag{5}$$

And finally, we can solve for x :

$$x = \frac{B(x) * width}{2 * \tan\alpha * B(z)} \tag{6}$$

The calculation of y is done equivalently. Instead of the width, the height is used and the horizontal view angle is replaced by the vertical view angle.

We repeat this procedure for every point in the object's outline and connect the points in the projection following the input sequence. Hence, the polygon can be displayed on the screen.

3.3 Sensor Fusion

Determining geographical locations requires sensor data received from GPS and orientation sensors of current smartphones. As orientation and GPS sensors don't provide very accurate data due to hardware and environmental factors (e.g., reflections) the information needs to be filtered. We implement different methods for sensor fusion and noise elimination. E.g., we weight the incoming GPS readings according to their timestamp, as typically more recent information provides more accurate information. We smooth the positioning information by calculating the average of this weighted value and previous weighted values.

4 Evaluation

With MapAR we want to explore the possibilities and limits of AR in geographic application scenarios. We designed different test cases under different conditions. We have chosen areas in the real-world under controlled and varying conditions and evaluated the projected areas with respect to accuracy of area, angles, perimeter and distance.

4.1 Evaluation Setup

For testing the precision of projecting objects under markerless everyday conditions with consumer devices, we decided to project parking lots as reference objects, as they have a defined rectangular shape of the size 5 x 2.35 meters and are visible on satellite imagery. With this simplistic shape we also can easily assess the properties of the projection with respect to the real-world object. The used device was a Samsung S3.

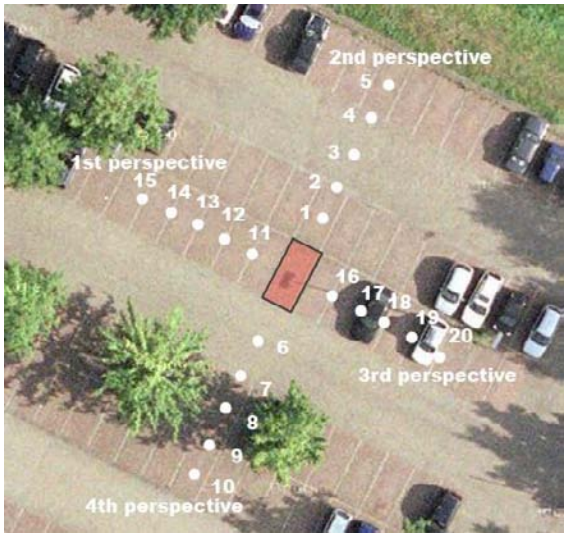
We recorded screenshots from projected parking lots. On a desktop computer with a 24" screen we manually selected the corner points of the projected rectangle with very high precision (we used a 27" screen with a resolution of 2560x1440 pixels, images where zoomed in to identify the correct position as precise as possible). We then translated the projection into geographical coordinates and reversely calculate the deviations from the correct parking lot, see Figures 4 and 5 for an illustration of the work flow.

In order to evaluate MapAR under realistic conditions we evaluated the result with four different variations (see Fig. 3):

- ***Differing perspectives***: we recorded 4 different perspectives for each parking spot in varying distances between 3 and 8 meters in order to rule out influences on perspective adaptation of the method.
- ***Differing distances***: we recorded each parking lot from 5 different distances (2.5, 5.0, 7.5, 10.0, 12.5 meters).
- ***Multiple recordings***: due to varying accuracy of GPS positioning we recorded two pictures for every position to rule out obvious outliers.
- ***Differing entities***: we used two different parking lots.

The conditions in our evaluation setup resulted in 80 individual measurements of the projection.

Figure 3: 5 x 2.35 meter parking spot in 4 different perspectives and different distances



.

Figure 4: The correct parking lot is outlined on the gound.

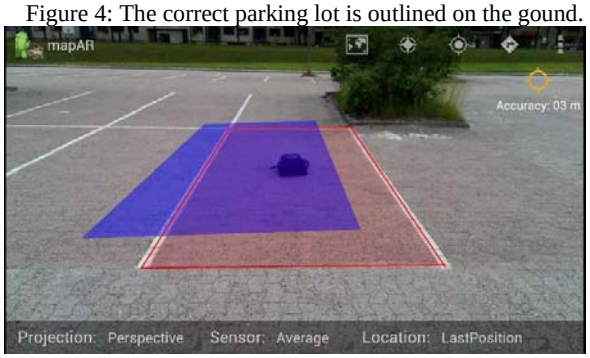


Figure 5: Translating the projection back to (geographic) world coordinates and projecting them back on the used satellite imagery.

We then compared the resulting 80 projected polygons with the original source polygon with respect to following properties:

- ***Center point distance***: the distance from the center of the projected polygon to the correct polygon (positioning accuracy).
- **Area**: we compared the area of the projected polygon with the correct polygon.

- ***Interior angles***: since the parking space is a rectangle, each interior angle has to be 90 °. We measure the deviation of the interior angles of the projected polygon.
- ***Perimeter***: Each parking lot has a perimeter of 14. 70 m. The perimeter of the projected polygon is compared to this value.

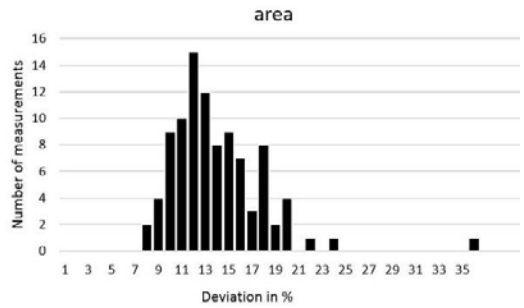
4.2 Results

The deviation of the distance to the center point of the parking lots has two peaks. While the first peak (16% of measurements) expresses a comparable small deviation of below 2m, the second peak (80% of measurements) clearly shows a relatively high deviation of up to 6 meters. This is due to the current positioning accuracy achievable with consumer grade GPS sensors (see Figure 6).

Figure 6: Deviation of the distance of the parking lots

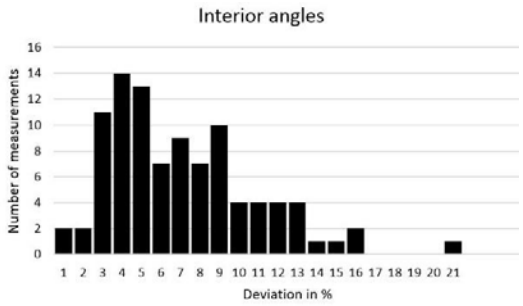
Our results show different deviations for our measurements. The deviation of the area of the projected parking lot is between 7 to 20 percent (Figure 7).

Figure 7: Deviation of the area of the parking lots



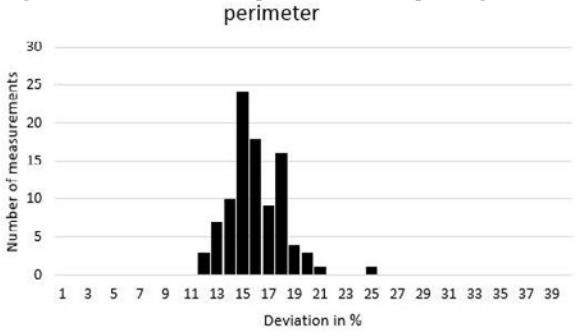
The deviation of interior angles has a peak between 4%- 9% indicating that the rectangle shape of the parking lot is well sustained in the projection (Figure 8).

Figure 8: Deviation of interior angles of the parking lots



The deviation of perimeter has a peak between 13% - 19% with almost 90% of all measurements inside of it (Figure 9). This also indicates a good maintenance of shape and size.

Figure 9: Deviation of the perimeter of the parking lots



5 Discussion

When interpreting the obtained results of our evaluation by means of geographic entities, we can identify the fields of application of markerless AR within geographic applications.

- ***Center point distance***: A large number of measurements (80,2%) showed a distance deviation of 5-6m, due to GPS inaccuracy. Typical entities of this dimension are smaller streets, smaller buildings, larger cars, parking lots, footprints of individual trees, etc. Any object of these or similar classes, depending on the configuration, might not be precisely addressable: if a similar entity is located directly next to the one to be augmented, in many cases the wrong entity will be augmented. I.e., if the entities are of a size in the range of the deviation augmentation is advisable only if the distance is large enough to guarantee disambiguation.
- **Area**: Although areal deviation is also in a perceivable range, most applications will still make sense, as large deviations in distance and shape might in many cases be more problematic. Many projected entities will have a certain counterpart in the real world and will be possible to correctly identify this entity even if the correct size is not preserved. As not arbitrarily large

entities can be projected to full extent, the achieved accuracy will often be below the distance error.

- ***Interior angles***: Our evaluation shows that geometry is preserved to a very high degree, indicating that information of sensors of the device itself already precise allows precise projections (within geographic application context).
- ***Perimeter***: 87% of measurements are between 13% - 18% deviation. This result is similar to the area deviation.

In the current state of technology (which is mainly limited by positioning accuracy), AR applications are applicable for entities of the size of the positional deviation or above. If the entity is perceivable without the help of augmentation and is a rather unique entity with respect to its surrounding, it can be also smaller.

I.e., in scenarios where precision (of currently) <5m is not required or entities can be perceived and matched due to their physical properties, it is feasible to use AR techniques in conjunction with consumer technology. However, in many cases this excludes scenarios without visually perceivable entities: examples are underground infrastructural elements like pipes, cables, or small scale excavation sites; identifying the correct entity can cause large efforts and costs.

The more alternative positioning systems (e.g., GLONASS, BeiDou, Galileo) and precision enhancing techniques are on the rise in the consumer market, the more can markerless AR be applied in geospatial high precision contexts with out-of-the-box consumer technology.

6 Conclusions

In this work we evaluated the applicability of AR techniques within the context of geographic applications.

As our evaluation shows, the application scenarios are mainly limited by the accuracy of the current predominant GPS positioning. This excludes a number of application scenarios from using AR as suitable method for identifying invisible properties or specific entities. Nevertheless there are numerous possibilities in which the application can be used with fewer requirements in terms of precision.

Acknowledgements

We gratefully acknowledge funding granted by the German Research Foundation (DFG) via the Transregional Collaborative Research Center SFB/TR8 Spatial Cognition, as well as funding granted by the European Union via mSAFE, grant agreement no. FP7-PEOPLE-2011-IRSES 295269.

References

[1] F. Liarokapis, I. Greatbatch, D. Mountain, A. Gunesh, V. Brujic-Okretic, J. Raper. »Mobile Augmented Reality Techniques for GeoVisualisation«. In: Proceedings of the Ninth International Conference on Information Visualisation. IV '05. Washington, DC, USA: IEEE Computer Society, 2005, S.745-751.

[2] F. Schmid, L. Frommberger, C. Chunyuan, C. Freksa. »What You See is What You Map: Geometry-Preserving Micro-Mapping for Smaller Geographic Objects with mapIT«. In: Geographic Information Science at the Heart of Europe. Eds. Danny Vandenbroucke, Bénédicte Bucher, Joep Crompvoets. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, 2013, S. 3–19

[3] R.Behringer. »Registration for outdoor augmented reality applications using computer vision techniques and hybrid sensors«. In: Virtual Reality, 1999. Proceedings., IEEE. 1999, S. 244–251.

[4] D. Stricker, T. Kettenbach. »Real-time and markerless vision-based tracking for outdoor augmented reality applications«. In: Augmented Reality, 20001. Proceedings. IEEE and ACM International Symposium on. 2001, S. 189–190.

[5] R. Azuma, B. Hoff, I. Neely H., R. Sarfaty. »A motion-stabilized outdoor augmented reality system«. In: Virtual Reality, 1999. Proceedings., IEEE. 1999, S. 252–259.

[6] Y. Wu, M. E. Choubassi, I. Kozintsev. »Augmenting 3D urban environment using mobile devices«. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on. 2011, S. 241–242.

[7] G. W Roberts, A. Evans, A. Dodson, B. Denby, S. Cooper, R. Hollands u. a. »The use of augmented reality, GPS and INS for subsurface data visualization«. In: FIG XXII International Congress. 2002, S. 1–12.

[8] A. H. Behzadan, V. R. Kamat. »Visualization of construction graphics in outdoor augmented reality«. In: Proceedings of the 37th conference on Winter simulation. WSC '05. Orlando, Florida: Winter Simulation Conference, 2005, S. 1914–1920.

[9] E. Veas, R. Grasset, E. Kruijff, D. Schmalstieg. »Extended Overview Techniques for Outdoor Augmented Reality«. In: Visualization and Computer Graphics, IEEE Transactions on 18.4 (2012), S. 565–572.

[10] G. R. King, W. Piekarski, B. H. Thomas. »ARVino – Outdoor Augmented Reality visualisation of viticulture GIS data«. In: Proceedings of the forth IEEE and ACM International conference on Mixed and Augmented Reality (ISMAR), oct 5-8, Wienna. 2005, S. 52–55

[11] B. Avery, W. Piekarski, B. H. Thomas. »Visualizing Occluded Physical Objects in Unfamiliar Outdoor Augmented Reality Environments«. In: In 6th Int'l Symposium on Mixed and Augmented Reality. p. 2007, S. 285–286

[12] Webster, S. Feiner, B. Macintyre, W. Massie, T. Krueger. »Augmented reality in architectural construction, inspection and renovation«. In: Proc. ASCE Third Congress on Computing in Civil Engineering. 1996, 913–919

[13] H. Goldstein. »Classical Mechanics«. 2nd. Reading, MA: Addison-Wesley Publishing Company, 1980, 146–148.

[14] K.K.F. Riley, M.P. Hobson, S.S.J. Bence. »Mathematical methods for physics and engineering«. Cambridge University Press, 2006, S. 931-942.

[16] F. Schmid, L. Frommberger, C. Cai, and F. Dylla. 2013. »Lowering the barrier: how the what-you-see-is-what-you-map paradigm enables people to contribute volunteered geographic information«. In Proceedings of the 4th Annual Symposium on Computing for Development (ACM DEV-4 '13). ACM, NY, USA

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Multi-sensory Integration for a Digital Earth Nervous System

Frank Ostermann
University of Twente – ITC, Faculty of Geo-Information Science and Earth Observation (ITC)
P.O. Box 217
7500 AE Enschede, The Netherlands
f.o.ostermann@utwente.nl

Sven Schade
European Commission – DG Joint Research Centre (JRC)
Via E. Fermi 2749
21027 Ispra, Italy
sven.schade@jrc.ec.europa.eu

Abstract

The amount of geospatial data is increasing, but interoperability issues hinder integrated discovery, view and analysis. This paper suggests an illustrative and extensible solution to some of the underlying challenges, by extending a previously suggested Digital Earth Nervous System with multi-sensory integration capacities. In doing so, it proposes the combination of multiple ways of sensing our environment with a memory for storing relevant data sets and integration methods for extracting valuable information out of the rich inputs. Potential building blocks for the implementation of such an advanced nervous system are sketched and briefly analysed. The paper stimulates more detailed considerations by concluding with challenges for future research and requesting a multidisciplinary development approach – including computer sciences, environmental sciences, cognitive and neurosciences, as well as engineering.

Keywords: interoperability, sensing, observation, multi-sensory integration, digital earth.

1 Introduction and motivation

The increasing amount of geospatial data that is available from new and existing sources has inspired numerous businesses, (non-)governmental initiatives and research projects to explore ways to utilize it. The heterogeneity of data sources and diverse processing histories imply issues of syntactic and semantic interoperability. Hence, many research initiatives and projects aim to improve data interoperability. Many tackle the problem with a bottom-up approach by developing proprietary solutions for specific business problems (e.g. Xively[1], Gigwalk[2], Jana[3]), or by developing open-source solutions that allow syntactical (e.g. GDAL[4], Web 2.0 Broker[5]), or semantical (e.g. HALE[6]) translation between concrete data sources, formats and standards. Most of these have a decidedly technical perspective on standards for data formats and data exchange protocols. Others approaches address the problem top-down and aim to develop new standards that facilitate discovery, view and analysis of heterogeneous data sources. The resulting standards address interoperability on a technical level (e.g. OGC[7], ISO[8], [9]), on a semantic level (e.g. common vocabularies and code lists, e.g. DublinCore[9]), but also on a governance and legal level (INSPIRE[10], ISA[11]).

These two perspectives have resulted in substantial advances in science and operational systems. Still, all these efforts face the problem of ensuring interoperability among themselves. It is already difficult to keep track of the past and ongoing efforts, let alone to coordinate them. Although mostly adhering to common data exchange standards, the projects and initiatives originate from various academic, administrative or entrepreneurial backgrounds, and thus do not always share ideas of and approaches to interoperability. Furthermore, while opening existing data silos in formerly closed spatial data infrastructures (SDI), new silos are created as part of the process - both vertically (e.g. through incompatible organizations), and horizontally (e.g. through incompatible service buses or middleware).

The interoperability issue is aggravated by the fast-moving technological landscape: (1) new opportunities (read: platforms) emerge quickly, while others are abandoned (e.g. Gowalla[12]) or face an uncertain future (e.g. Foursquare[13]); (2) many web portals are no longer maintained after funding stopped, but many diverse government portals offers data [3]; (3) out of the numerous citizen science projects (see Sci-Starter[14] and Zooniverse[15] platforms and JRC Citizen Science and Smart Cities 2014 Summit[16]), many come with proprietary software applications; and (4) initiatives such as INSPIRE move slowly because of the legislative requirements and number of partners involved, and have difficulty adapting

[1] https://xively.com/
[2] http://gigwalk.com/
[3] http://www.jana.com/
[4] http://www.gdal.org/
[5] http://www.geotec.uji.es/web-2-0-broker-service/
[6] http://www.esdi-community.eu/projects/show/hale
[7] http://www.opengeospatial.org/
[8] http://www.isotc211.org/
[9] http://dublincore.org/
[10] http://inspire.jrc.ec.europa.eu/
[11] http://ec.europa.eu/isa/
[12] http://blog.gowalla.com/
[13] http://www.foursquare.com/
[14] http://scistarter.com/
[15] https://www.zooniverse.org/
[16] http://ies.jrc.ec.europa.eu/DE/derdu-latest-news/sdi-workshops/citizens-science-and-smart-cities-summit.html

to new technological developments, e.g. linked open data (for a discussion of differences, see Portele, C.[17]).

This paper offers an original perspective on the problem outlined above by extending and revising the conceptual model of a Digital Earth Nervous System (DENS) with the process of Multi-Sensory Integration (MSI), drawing on rich research from the cognitive and neuro-sciences, as well as sensor data fusion from engineering. The aims are threefold: (i) to stimulate and enrich the debate on interoperability for geospatial data; (ii) to increase understanding of the various interactions between geospatial data collection, transformation, processing and usage on a global scale; and (iii) to show potential future research foci. The DENS-MSI should be able to serve as a possible reference and orientation for existing approaches and projects to increase mutual understanding of interoperability challenges and how to deal with conflicting information in a decision-making environment.

The paper is not trying to create a conceptual or logical model which is complete (and overly complex) and suitable for every circumstance and situation possible. Instead it focusses on in-situ sensory and citizens' observations and aims to be simple, extensible (open world assumption), and cover the majority of cases. Neither is it meant to promote a 21st century version of the Gaia hypothesis, from which the authors would like to distance themselves.

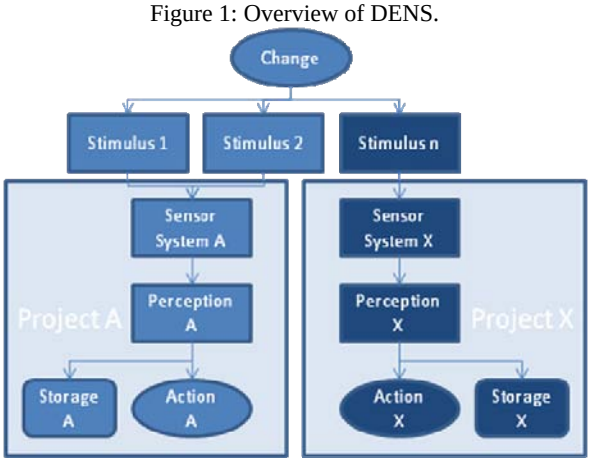
In the next section, this paper gives a short introduction to and critique of the original Digital Earth Nervous System, and its reception and usage since then. The section following it briefly explains the background of the MSI concept, which is one focus of this paper's extension of the previously suggested DENS. The last section of the paper sketches a possible integration of the DENS and MSI, and paths for future research.

2 A Digital Earth Nervous System

The DENS concept was originally formulated by DeLongueville et al. [5]. It draws an analogy to the human nervous system in order to describe and understand the processing of inputs from geospatial sensors (compare Figure 1). Here, many types of digital data and information with a geographic component can form sensory input (stimuli in Figure 1), i.e. the sensory input can range from remotely sensed spectral information of the earth's surface to geo-locatable text messages that are exchanged between citizens.

The great strength of this approach lies in its unifying vision of treating all geospatial information as potential input. It acknowledges the rise of volunteered geographic information [6, 8] and sensor networks of cheap and wireless hardware (e.g. Zigbee[18], Raspberry Pi[19]), and the need for utilizing it together with authoritative data from SDIs (see SDI cookbook chapter 10[20]), e.g. as part of quality assurance procedures. It also provides suggestions for methods to collect and store this heterogeneous geospatial information, focusing on the OGC Sensor Web Enablement (SWE) standard [2].

Figure 1: Overview of DENS.



Source: The authors.

Several studies have drawn on or from the DENS concept, e.g. a functional integration approach for the sensor web [18] and a way to sense VGI for disaster management [19]. These studies show that the DENS concept offers a valuable perspective to create original and successful ways to interact and use the various information provided. It is a reasonable assumption that developments such as cloud computing[21] and linked open data [1, 4] will improve feasibility of a DENS implementation.

However, some of the studies also showed that the DENS analogy is not suitable for all cases, or the data cannot be clearly assigned to every phase. For example, not all detailed phases of sensor processing proposed in [5] were applicable in [16]. Further, the SWE suite of standards is rather complex to implement and will not be the method of choice for many potential VGI sources – although lightweight RESTful implementations are in development [13].

The envisioned treatment of the uncertainty of VGI is another shortcoming. DeLongueville et al. [5] originally suggest that VGI needs to be validated before it is made available as an observation, but do not propose possible implementations. The current DENS approach cannot explain conflicting sensor inputs, e.g. the presence of Tweets about forest fires in an area for which remote sensing does not indicate any hot spots [20]. The human cognitive system has developed methods to deal with conflicting multi-sensory input. Contradictory sensor input can be resolved at the level of raw sensor data (stimuli and sensations) in order to check for obvious errors in sensor readings with the potential result of a re-calibration. An alternate opportunity addresses conflicts at the level of perceptions, potentially resulting in the re-evaluation of a perception.

For the latter, Spinsanti and Ostermann [20] successfully adapted an argument by Flanagin and Metzger [7] on the heuristics that humans use to deal with uncertain information: by looking into other sources ("What do others say?") and comparing the new information with existing knowledge

17 http://www.pilod.nl/index.php?title=Boek/Portele#Technical_Comparison_of_Linked_Data_and_INSPIRE

18 https://www.zigbee.org/

19 http://www.raspberrypi.org/

20 http://www.gsdidocs.org/GSDIWiki/index.php/Chapter_10

21 http://www.nist.gov/itl/cloud/

("What do I already know?"). However, the resulting method (GeoCONAVI) shows that currently it is computationally most expensive in the early stages to reduce noise, yet the most significant improvement on information quality occurs at the later stages of the processing chain, when the information had already been consolidated and clustered [16]. Multi-Sensory Integration might provide a solution for an early validation and treatment of inconsistent sensor input. We explore this option in the next section.

3 Multi-Sensory Integration

Multi-sensory integration (MSI) – also known as multi-modal integration – encompasses the process of combining the information from different sensory systems, such as visual, audio, tactile, olfactory, taste and interoception[22] by the nervous system. It is thus a crucial process without which there would be no coherent representation of the environment, and no interpretable perceptual experience. Therefore, it is also the prerequisite for any adaptive behavior and response to the environment. An important aspect of human MSI is the mutual feedback between sensory systems. Research has shown that for example visual and auditive systems influence each other, i.e. a strong signal on one "channel" can alter the perception of the other.

The nervous system integrates or segregates groups of sensory signals based on three major principles of multi-sensory integration: spatial proximity, temporal proximity, and inverse effectiveness. The first two are analogous to Tobler's First Law, while the inverse effectiveness supports an assumption that is present in the work of Spinsanti and Ostermann [20], i.e. that multiple sensor readings from different but weak sensors can together result in a valid and coherent perception. Thus, MSI results in decreased sensory uncertainty. Another desirable effect are decreased reaction times – while a system might need many stimuli from just one sensor, fewer stimuli from many sensors can lead to the same conclusion.

There are several approaches to explaining human MSI, such as visual dominance, modality appropriateness, and Bayesian integration [21]. Especially the latter might integrate well with spatio-temporal data handling. A major challenge for Bayesian integration is the assignment of probabilities of conditions to observed stimuli.

In the field of sensor engineering, the research area of sensor data or information fusion has already seen a lot of activity [14]. The majority of research until now has focused on low-level abstracted sensor data, i.e. low-dimensional, continuous data from sensors with a known uncertainty, on data fusion from several but similar sensors, or on different but related sensors in close spatial proximity (e.g. robotics). The integration of heterogeneous sensors covering irregular areas, e.g. wireless sensor networks from citizens or geosocial network data (hard/soft data integration from disparate sensors in the terminology of [14]) has seen less activity.

The following section will investigate how these concepts from cognitive science and information fusion could be fruitful motivations for future research in the areas of (geo)sensor web and (geo)social networks.

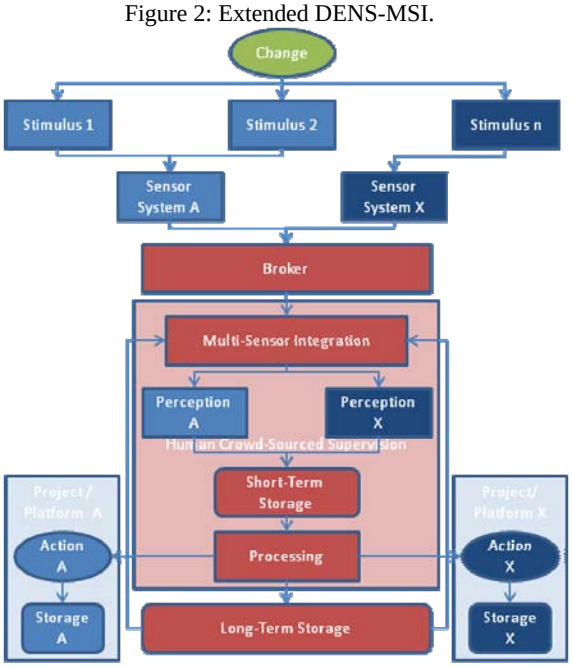
[22] sensitivity to stimuli originating inside of the body;

4 Design and implementation of a DENS

In this section, we show how concepts and theories from neuroscience and robotics can contribute to an overall understanding and improvement of geospatial data interoperability on a global scale. We directly build on previous work on DENS, which only addressed observations from a single source and sequences of data flows.

The following Figure 2 shows an extended and revised DENS-MSI and contains all the elements and processes we will discuss. As we will argue, this raises three main challenges: first, the choice of senses (sensors) and their interoperability; second, the choice of memory (geospatial data sets); and third, the choice of actual MSI methods.

Figure 2: Extended DENS-MSI.



Source: The authors

It all begins with an observable change in the environment, for example, in the case of a forest fire remote sensing, satellites can detect higher temperatures on the ground, smoke plumes, and citizens and practitioners on the ground begin to discuss and share information. This creates stimuli which are observed by sensor systems, e.g. Twitter, OSM, Flickr or satellites.

Considering the many potential sensors that the DENS can "listen" to, we need to identify those with the highest likelihood of containing information about the phenomenon that we are interested in (the right cues for combination). Thus, we need to have prior knowledge about the phenomenon and codify it in rules. For example, the utility of some sensor systems depend on the time of the day (just as human sensor systems do), e.g. whether it is day or night. As a first step, a brokering [15] approach[23] can help to integrate the sensor data on the technical and syntactic interoperability level. The next level would be semantic integration or

[23] http://www.essi-lab.eu/do/view/GIaxe/WebHome

interoperability through ontologies, metadata, and vocabularies [17].

Yet, it remains questionable whether it is feasible to semantically enrich sensor data on a low (atomic) level [10], because of the number of potential sensors readings that need processing and the exponential growth of links that might not in fact be sensible. It seems more appropriate to do the linking and semantic enrichment on the higher level of perceptions. On the level of individual stimuli, it seems more reasonable to check (i) whether the source is trusted (or neutral); and (ii) which (if any) detectable keywords are included, instead of analyzing the content and context in detail. The resulting sensor set can then be used for the actual MSI.

The three major principles from neuroscience and cognitive science (spatial and temporal proximity, inverse effectiveness) show a clear alignment with the core principles of processing spatio-temporal data: what is near in space and time is related. This strengthens the analogy between human and digital earth nervous system. If multiple sensory inputs are available, then a DENS-MSI can rely on cue combination, i.e. a comparison of the various sensor inputs. In the optimal case, these can be unified in single coherent perception (e.g. remote sensing shows smoke plume over forests, Tweets talk about fire). However, if the cues are dissonant (e.g. Tweets show talk about forest fire in location X, but a visual live stream from a web cam showing nothing extraordinary or no smoke), causal inference provides an alternative. Causal inference is a crucial component in human perception and uses prior knowledge to resolve the conflicting sensor inputs by resorting to causal structures. It determines the most plausible causal structure to explain the dissonance. In the example above, possible causal structures are a sensor misreading (interpretation of Tweets), or temporal misalignment (remote sensing images do not match the exact same period). Here we tap into cognitive processes, especially long-term memory retrieval, in order to determine the most likely causal structure. For the MSI, the system would have to be able to assign likelihoods based on prior knowledge codified as machine-readable information. This is analogous to the GeoCONAVI use of authoritative datasets [16]. Given the large number of data sets available, we need to identify those that are the most relevant for the task or phenomenon. This corresponds to geographic information retrieval, with the important question: which data sets (i.e. knowledge) to choose? Ivanova [12] explores a solution based on domain expert input.

The research from sensor data fusion has only recently begun to investigate the particular issues found with integration disparate sensors and hard/soft data, i.e. geospatial sensor networks from humans, low-cost in-situ sensors, and remote sensing. However, in addition to the Bayesian probabilistics discussed above, possibilistic and human centered approaches are investigated. While the former offers potential solutions that need further exploration, the latter one relates to crowd-sourcing tasks (see below).

Continuing our thought experiment, the integrated sensory information results in perceptions of events on the Earth, e.g. forest fires. We can expect many such perceptions. These and the corresponding stimuli are stored in a short-term memory for immediate reference. This short-term memory is constantly analyzed (searched for patterns) and monitored. Only when a number of criteria (rules) are fulfilled is an alert being raised (e.g. several perceptions relating to forest fires in close spatial and temporal proximity). Similarly to the MSI, this filtering can be supervised by crowd-sourcing efforts.

As a last step, verified sensor information can be stored in a long-term memory to be accessed for future multi-sensory integration, or other geographic information retrieval tasks. The short-term memory and long-term memory together form a 'Digital Earth Memory System'.

Clearly, a challenge is to train such a semi-autonomous system to filter and sort stimuli, query existing data sets for validation, integrate heterogeneous sensor data and monitor perceptions that are stored. Supervised machine learning would need constant human supervision, but this is actually a process that can be very well crowd-sourced. A constant stream of a stratified sample of the DENS perceptions could be used for this purpose. The stimuli that are part of these perceptions are checked by volunteers and micro-tasked paid crowd-workers. Hung [11] shows the feasibility of methods to filter out spammers and low-quality contributions. For example, they could check whether a Tweet that supposedly belongs to a perception "forest fire near Avignon, France" is actually about a forest fire in France). Gamification offers even more opportunities, e.g.[15].

5 Conclusions and outlook

This paper aimed to stimulate the debate on interoperability for geospatial observations, to increase understanding of the various interactions between geospatial data collection, transformation, processing and usage on a global scale, and to show potential future research foci.

Indeed, the paper has highlighted developments in and important challenges for improving interoperability of heterogeneous geospatial data sources. We have argued that the concept of the DENS can help and improve mutual understanding between practitioners, researchers, developers and citizens. Further, the paper has shown how knowledge from the disciplines of cognitive and neurosciences, as well as engineering can contribute to an improved DENS model.

Particularly promising research objectives include the assessment of a sensor's observations' validity through possibilistic methods and the use of crowd-sourcing to supervise machine learning of algorithms and rules to filter, sort and organized stimuli into coherent perceptions.

Arguably, too specific approaches had little success in increasing interoperability until now, while there is some risk of failure for over-generic approaches. Therefore, we suggest following a stepwise and incremental development methodology. We plan to use well examined cases, such as the forest fire [16, 18, 20] or flood [19] examples for the initial set-up of a possible solution, before moving into new areas. Here, we will address urban environments, which should provide a solid ground for, especially because of the related high traffic in social media.

References

[1] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *5*(3), 1–22.

[2] Bröring, A., Echterhoff, J., Jirka, S., Simonis, I., Everding, T., Stasch, C., Lemmens, R. (2011). New Generation Sensor Web Enablement. *Sensors*, *11*(3), 2652–2699.

[3] Dukaczewski, D., Ciolkosz-Styk, A., & Sochacki, M. (2013). Regional geoportals of first-level administrative units of European Union and European Economic Area countries - Comparative Study. In M. F. Buchroithner (Ed.), *Proceedings of the 26th International Cartographic Conference*. Dresden: International Cartographic Association.

[4] Freitas, A., Curry, E., Oliveira, J. G., & O'Riain, S. (2012, October 17). Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends, *16*(1), 24–33.

[5] De Longueville, B., Annoni, A., Schade, S., Ostlaender, N., & Whitmore, C. (2010). Digital Earth's Nervous System for crisis events: real-time Sensor Web Enablement of Volunteered Geographic Information. *International Journal of Digital Earth*, *3*(3), 242 – 259.

[6] Elwood, S., Goodchild, M. F., & Sui, D. Z. (2011). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590.

[7] Flanagin, A., & Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3), 137–148.

[8] Goodchild, M. F. (2007). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, *2*(1), 24–32.

[9] Havlik, D., Schade, S., Sabeur, Z. A., Mazzetti, P., Watson, K., Berre, A. J., & Mon, J. L. (2011). From Sensor to Observation Web with Environmental Enablers in the Future Internet. *Sensors*, *11*(4), 3874–3907.

[10] Henson, C. A., Pschorr, J. K., Sheth, A. P., & Thirunarayan, K. (2009). SemSOS: Semantic sensor Observation Service. In *Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems* (pp. 44–53). IEEE Computer Society.

[11] Hung, N. Q. V., Tam, N. T., Tran, L. N., & Aberer, K. (2013). An Evaluation of Aggregation Techniques in Crowdsourcing. In *Proceedings of WISE 2013* (Vol. 8181, pp. 1–15). Springer.

[12] Ivánová, I., Morales, J., de By, R. A., Beshe, T. S., & Gebresilassie, M. A. (2013). Searching for spatial data resources by fitness for use. *Journal of Spatial Science*, *58*(1), 15–28.

[13] Janowicz, K., Bröring, A., Stasch, C., Schade, S., Everding, T., & Llaves, A. (2011). A RESTful proxy and data model for linked sensor data. *International Journal of Digital Earth*, *6*(3), 233–254.

[14] Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, *14*(1), 28–44.

[15] Nativi, S., Craglia, M., & Pearlman, J. (2012). The Brokering Approach for Multidisciplinary Interoperability: A Position Paper. *International Journal of Spatial Data Infrastructures Research*, *7*, 1–15.

[16] Ostermann, F., & Spinsanti, L. (2012). Context Analysis of Volunteered Geographic Information from Social Media Networks to Support Disaster Management: A Case Study On Forest Fires. *International Journal of Information Systems for Crisis Response and Management*, *4*(4), 16–37.

[17] Perego, A., Fugazza, C., Vaccari, L., Lutz, M., Smits, P., Kanellopoulos, I., & Schade, S. (2012). Harmonization and Interoperability of EU Environmental Information and Services. *Intelligent Systems, IEEE*, *27*(3), 33–39.

[18] Schade, S., Ostermann, F., Spinsanti, L., & Kuhn, W. (2012). Semantic Observation Integration. *Future Internet*, *4*(3), 807–829.

[19] Schade, S., Díaz, L., Ostermann, F., Spinsanti, L., Luraschi, G., Cox, S., De Longueville, B. (2013). Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information. *Applied Geomatics*, *5*(1), 3–18.

[20] Spinsanti, L., & Ostermann, F. (2013). Automated geographic context analysis for volunteered information. *Applied Geography*, *43*(0), 36–44.

[21] Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22–39.

Session:
Land Cover and Use
COST TD1202

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database

Jacinto Estima
ISEGI – Universidade Nova de Lisboa
Campus de Campolide
Lisboa, Portugal
Jacinto.estima@gmail.com

Cidália C. Fonte
Department of Mathematics
University of Coimbra /
INESC Coimbra
Coimbra, Portugal
cfonte@mat.uc.pt

Marco Painho
ISEGI – Universidade Nova de Lisboa
Campus de Campolide
Lisboa, Portugal
painho@isegi.unl.pt

Abstract

Volunteered Geographic Information has been increasing exponentially over the last years, capturing the attention of the scientific community. Researchers have been very active exploring a vast amount of initiatives and trying to develop methodologies and possible real applications for this new source of geographic information. Land Use/Cover production is one of the areas where this type of geographic information might be very useful. In this paper we evaluate if geo-referenced and publicly available photos from the Flickr initiative can be used as a source of geographic information to help Land Use/Cover classification. Using the Corine Land Cover nomenclature, we compare the classification obtained for selected photo locations, against the classification obtained from high resolution satellite imagery for the same locations. We conclude that this source cannot be used alone for the purpose of Land Use/Cover classification but we also believe that it might contain helpful information if combined with other sources.

Keywords: Volunteered Geographic Information, Land Use/Cover, Flickr, Accuracy assessment.

1 Introduction

The availability of Volunteered Geographic Information (VGI), a term coined by Michael Goodchild in 2007 [1], has been increasing exponentially in the last years, due to the introduction of the Web 2.0 and the increasing availability of low cost positioning equipment, among other technology improvements [2, 3].

The scientific community has been trying to explore this vast amount of information to use it in the solution of world problems. Navigation [5], crisis management and emergency response [6, 7] are just a few examples of the research that has been conducted in the last years. Although some of the main issues and concerns pointed out about this type of Geographic Information (GI) are related with their heterogeneity, quality control and metadata absence, among others, it is also agreed that their main advantages are associated with their temporal coverage and volume [8].

More recently, Fischer [4] argued that, in some cases, when VGI is used for different purposes than those for which volunteers have contributed, it can be seen as a not-so-Volunteered Geographic Information and had termed this as involuntary geographic information (iVGI).

The specific application of this type of data for Land Use/Cover (LULC) production had also been investigated recently, achieving some interesting results (e.g. [9, 10]). For example, Urban Land Use was produced using data from OpenStreetMap [11]. Photos and descriptions extracted from the Degrees of Confluence Project were used to assess the accuracy of Land Cover Maps [12, 13]. In a paper studying Flickr photos [9], the authors explored a collection of Flickr geotagged and publicly available photos in terms of their temporal and spatial distributions over Continental Portugal and also its distribution over Land Use/Cover classes, using as a reference the European Corine Land Cover (CLC) database. They concluded that this source of VGI might be very valuable for the purpose of LULC production when combined with other sources.

In this paper we evaluate whether what is seen on Flickr photos can be used to provide a LULC class or not and, whenever it is possible to do so, we evaluate if the identified class is correct. This evaluation was made using a stratified sample of photos considering the CLC level 1 classes as strata, and comparing the classes extracted from the Flickr photos to the class assigned to the photos location using high resolution satellite imagery.

The paper is structured as follows. After the introduction we describe the methodology used, followed by the presentation and discussion of the obtained results. The paper ends drawing some conclusions and indicating future research directions.

2 Material and Methods

2.1 Description of the study area and datasets

The defined study area is the Portuguese municipality of Coimbra, covering an area of approximately 300 km^2.

Three datasets were used in this study: 1) the geo-referenced Flickr photos for the study area over the period ranging between 2004 and 2013, corresponding to a total of 4977 photos; 2) the CLC database, composed by the version 16 (04/2012) for the CLC2006 inventory, downloaded from the

European Environment Agency (EEA)[1]; 3) the high resolution satellite imagery, with 30cm spatial resolution, available for the study area at the ArcGIS software as basemap.

Figure 1shows the CLC map for the study area and Figure 2 shows the points corresponding to the spatial location of the photos situated in each of the three CLC classes used for this analysis, overlaid with the high resolution satellite images.

Figure 1 - CLC level 1 classes in Coimbra municipality

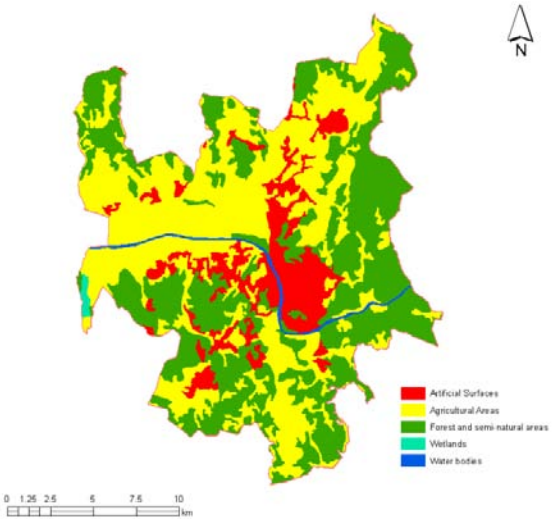
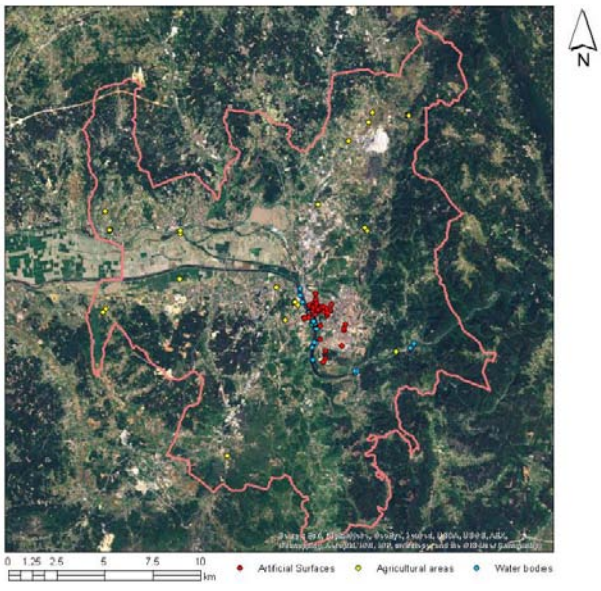


Figure 2 - Location of the sample Flickr photos used for the analysis



2.2 Data processing

For the purpose of this preliminary study, the position associated to the 4977 photos was intersected with the CLC level 1 classes and the three classes that from a user perspective were more likely to have information were selected, namely classes 1, 2 and 5, respectively Artificial Surfaces (AS), Agricultural Areas (AA) and Water Bodies (WB), corresponding to a total of 4892 photos.

Table 1 summarizes the distribution of selected photos over the three CLC classes.

Table 1 - Summary of Flickr photos

CLC Classes	Flickr Photos
Class AS	4703
Class AA	64
Class WB	125
Total	4892

2.3 Methods

The methodology adopted to conduct this analysis was as follows:

1. A stratified sample of 60 photo locations was selected for each of the three classes chosen for the analysis, considering the CLC classes as strata;
2. An expert classification of Flickr photos was done, based on the image content interpretation, according to the CLC nomenclature. Using the photo assigned to each location, we first evaluate whether it was possible to attribute a class or not and, when possible, a class was then assigned to the corresponding location;
3. Flickr photo locations were overlaid with the high resolution satellite imagery, and a land cover class was assigned to each location based on the imagery interpretation.

3 Results and discussion

Following the methodology described in section 2.3, Table 2 and Table 3 show the resultant classification of the locations based on the interpretation of Flickr photos and satellite imagery respectively. Besides CLC level 1 classes, two more classes were considered: “Not Clear” and “Not Good”. The “Not clear” class refers to those photos where more than one class is present and it is not clear which one, if any, is predominant, and the “Not Good” class refers to those photos that do not show predominantly any type of landscape and therefore cannot be used in LULC classification.

Having a closer look to the spatial distribution of the photos relatively to the CLC classes (see Figure 1 and Figure 2) it is clear that for class AS they are centered in the more touristic places of the city of Coimbra and for WB most photos are located in the region of the river where touristic boats operate. A more even distribution can be seen for the class AA.

Results from the interpretation of Flickr photos are shown in Table 2. The percentage of photos considered “not good” for LULC classification is relatively high, with 41.7% for class AS, 26.7% for class AA and 21.7% for class WB. Another negative aspect is related with the percentage of photos classified as “not clear”, with 18.3%, 26.7% and 6.7% for classes AS, AA and WB respectively. These two classes together, representing photos that do not fit in any CLC class, embody a high percentage of photos with classes AS, AA and

[1] http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-2

WB getting respectively 60%, 53.4% and 28.4%. In the opposite direction, the value for locations correctly classified is very low for all the classes with the class AA getting the worst value, below 20%. Looking at the value for photos wrongly classified, we can see a good result for class AS, with 0%, while classes AA had 28.3% and class WB 15%.

Table 2 - Classification of Flickr photos

		CLC Classes containing the photo's location		
		Class AS	Class AA	Class WB
Classification based on Flickr photos	Class AS	24	10	6
	Class AA	--	11	--
	Class F	--	--	3
	Class W	--	6	--
	Class WB	--	1	34
	Not Clear	11	16	4
	Not Good	25	16	13
Total of photos		60	60	60
Correct		40.0%	18.3%	56.7%
Wrong		0.0%	28.3%	15.0%
Not clear		18.3%	26.7%	6.7%
Not good		41.7%	26.7%	21.7%

Table 3 - Classification of Flickr photos' locations based on the satellite imagery

		CLC Classes containing the photo's location		
		Class AS	Class AA	Class WB
Classification based on satellite imagery	Class AS	60	1	--
	Class AA	--	39	--
	Class F	--	--	7
	Class W	--	--	--
	Class WB	--	--	52
	Not Clear	--	20	1
Total of points		60	60	60
Correct		100.0%	65.0%	86.7%
Wrong		0.0%	1.7%	11.7%
Not clear		0.0%	33.3%	1.7%

During the classification process, however, some problems related to the use of the Flickr photos became apparent, contributing to increase the negative aspects of this source. Among the collection of photos analyzed, we have seen photos showing predominantly people, photos taken inside houses, photos showing small details and photos taken far from what is shown in the image reflecting a high level of zoom. This last case was particularly present for photos considered inside class WS, where although the picture shows mainly water it is easy to realize that the pictures were taken from land.

The assignment of classes to the photo locations using the satellite imagery produced the results shown in Table 3. It can be seen that 100% of the points located at the AS areas in the CLC map where actually assigned to the class AS, with values of respectively 65% and 87% for the classes AA and WB.

At some locations it is not clear to which class the point should be assigned, due to the mixture of classes observed at the vicinity of the point and to the fact that the minimum mapping unit of the CLC map is 25ha, which means that the class choice cannot be done analyzing only what exists at each point, but also looking at a larger vicinity. In Table 3 it can be seen that this difficulty occurred for 20 points. However, a closer analysis showed that only 4 of these points correspond to different locations. The other 16 points, even though corresponding to different Flickr photos, actually were assigned exactly to the same spatial location, meaning that the volunteer assigned the same coordinates to a large number of photos. Moreover, an analysis of the photos as well as the photos tags also showed that there are also other photos wrongly geotagged, since the coordinates assigned are far from the real location were the photo was taken.

4 Conclusions and future research

The concentration of Flickr photos in touristic places, leading to a spatially poor distribution of locations, was confirmed. Consequently, urban areas and touristic places are likely to have more photos than rural areas. This spatial clustering needs to be assessed in order to understand its impact and the possibilities of its use to further parameterize the validation process. Besides, not all the photos were given a CLC class either because the predominant class was not clear or the image was not showing any type of landscape at all. As an example, for classes AS and AA, more than 50% of the photos could not be classified.

Some of the issues need further research in order to find possible solutions to overcome them. This is the particular case of WB class, where some pictures were taken from land and therefore their location is not inside the water body. One possible strategy to make those photos useful would be to buffer the water body with an acceptable distance and consider also inland locations inside that buffer.

Based on what has been exposed, we might conclude that this source of VGI is not suitable for LULC classification when used alone. Nevertheless we believe that it might contain useful information that can be helpful if combined with other sources. On the other side, different classes had different results and not all the classes were explored in this study. Therefore we are planning to continue this research to explore all the classes, compare the results at different locations and create pre-processing procedures that can improve the quality of the results.

It is also planned to combine this source of VGI with other sources, such as Panoramio, OpenStreetMap, among others, to understand if some of the faced issues would be solved and/or minimized when multiple sources are combined.

Acknowledgements: This research has been partially supported by COST Action TD1202, Mapping and the Citizen Sensor. C. Fonte work was partially supported by the Portuguese Foundation for Science and Technology under project grant PEst-OE/ EEI/UI308/2014.

References

[1] Goodchild, M. (Nov. 2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*. vol. 69, no. 4. pp. 211–221.

[2] Elwood, S., Goodchild, M. F., & Sui, D. Z. (May 2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Ann. Assoc. Am. Geogr*. vol. 102, no. 3. pp. 571–590.

[3] Heipke, C. (Nov. 2010). Crowdsourcing geospatial data. *ISPRS J. Photogramm. Remote Sens.* vol. 65, no. 6. pp. 550–557.

[4] Fischer, F. (2012). VGI as Big Data: A New but Delicate Geographic Data-Source. *Geoinformatics April/May*. no. May. pp. 46–47.

[5] Holone, H., Misund, G., & Holmstedt, H. (2007). Users Are Doing It For Themselves: Pedestrian Navigation With User Generated Content. in *International Conference on Next Generation Mobile Applications, Services and Technologies*. no. Ngmast.

[6] Goodchild, M., & Glennon, J. A. (Sep. 2010). Crowdsourcing geographic information for disaster response: a research frontier. *Int. J. Digit. Earth*. vol. 3, no. 3. pp. 231–241.

[7] Zook, M., Graham, M., Shelton, T., & Gorman, S. (Jan. 2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Med. Heal. Policy*. vol. 2, no. 2. pp. 6–32.

[8] Leung, D., & Newsam, S. (2010). Proximate sensing: Inferring what-is-where from georeferenced photo collections. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 2955–2962.

[9] Estima, J., & Painho, M. (2013). Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover. in *Computational Science and Its Applications – ICCSA 2013*. vol. 7974. pp. 205–220.

[10] Estima, J., & Painho, M. (2013). Exploratory analysis of OpenStreetMap for land use classification. in *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '13*. pp. 39–46.

[11] Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer, J., & Zipf, A. (2013). Toward mapping land-use patterns from volunteered geographic information. International Journal of Geographical Information Science, 27(12), 2264–2278.doi:10.1080/13658816.2013.800871

[12] Iwao, K., Nishida, K., Kinoshita, T., Yamagata, Y., 2006. Validating land cover maps with Degree Confluence Project information. Geophysical Research Letters 33.

[13] Foody, G.M., Boyd, D.S., 2013. Using Volunteered Data in Land Cover Map Validation: Mapping West African Forests. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 1305–1312.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Cropland Capture: A Gaming Approach to Improve Global Land Cover

[1]Linda See, [2]Tobias Sturn, [1]Christoph Perger, [1]Steffen Fritz, [1]Ian McCallum and [1]Carl Salk
[1]International Institute for Applied Systems Analysis
Schlossplatz 1
Laxenburg, A-2361, Austria
see@iiasa.ac.at

[2]Vienna University of Technology
Karlsplatz 13
Vienna, 1040, Austria
tobias.sturn@vol.at

Abstract

Accurate and reliable information on global cropland extent is needed for a number of applications, e g. to estimate potential yield losses in the wake of a drought or for assessing future scenarios of climate change on crop production. However, current global land cover and cropland products are not accurate enough for many of these applications. One way forward is to increase the amount of data that are used to create these maps as well as for validation purposes. One method for doing this is to involve citizens in the classification of satellite imagery as undertaken using the Geo-Wiki tool. This paper outlines Cropland Capture, which is simplified game version of Geo-Wiki in which players classify satellite imagery based on whether they can see evidence of cropland or not. On overview of the game is provided along with some initial results from the first 3 months of game play. The paper concludes with a discussion of the future steps in this research.

Keywords: Cropland, land cover, gaming, citizen science, crowdsourcing

1 Introduction

Accurate and reliable spatial information on cropland is essential for the estimation of potential yield losses that could occur as a result of wide spread drought or other anomalies that negatively affect crop production. Reliable cropland information is also needed for tackling other major environmental issues such as setting EU and US biofuel targets, determination of greenhouse gas emissions from different sectors including agriculture, REDD+ (Reducing Emissions from Forest Degradation and Deforestation) initiatives, and for determining the implications of climate change on crop production and patterns of productivity.

Global cropland extent can be obtained from global land cover products such as the GLC-2000 [10], MODIS [6], GlobCover [3] and the most recent 30m Chinese land cover product [17]. However, the problem with these products is that they are not accurate enough to provide a reliable estimate of croplands. For example, in Africa where there are extensive areas of low agricultural intensification, the spectral signatures and temporal profiles of cropland is similar to that of grasslands so differentiation between these two types is difficult [14]. Another issue is the lack of adequate data for training these maps using automated classification algorithms. The products also need further validation data. For this reason, the Geo-Wiki tool was developed, which is a visualization, crowdsourcing and validation tool for improving global land cover [8, 9]. Crowdsourcing [12], volunteered geographic information [11] and citizen science [2] are all terms for the involvement of citizens in data collection, analysis and scientific research of which Geo-Wiki is one of many applications.

In the past, a series of crowdsourcing campaigns were run to collect data using Geo-Wiki to help answer specific research questions regarding, e.g. land availability for biofuels [7], wilderness mapping [15] and land grabbing [1]. Although successful, we wanted to find methods for attracting larger numbers of participants and developing a much larger database for training and calibration. Gaming represents one potential way for achieving this. Games are currently the number one application used on smartphones [5], which represents an incredible number of potential players. Serious games, games with a purpose and gamification of existing applications are now becoming more common place [4, 13] so the idea of moving Geo-Wiki into a gaming environment was a logical step forward in encouraging citizens to participate. Previous serious games were tried with some success, e.g. [16], but we realised that a much simpler approach was needed.

This paper outlines the most recent development in the Geo-Wiki project, which is a game called Cropland Capture. The ultimate goal of the game is to gather training and validation data for improving global maps of cropland extent, which will be part of future research. The game is currently running and will end in May 2014. This paper outlines some initial results of the data gathered from the game and our plans for the future.

2 Cropland Capture

Cropland capture is a simple game in which players are presented with a red rectangle placed on top of satellite imagery from Google Earth. They are then asked to determine if there is any evidence of cropland in the image (Figure 1). They can answer yes, no or maybe if they are unsure. For each correct answer, the player receives a single point. For incorrect answers, the players lose one point. If a player answers maybe they do not gain or lose any points. We define correct answers in one of two ways. The first way involves expert intervention, where some of the pixels are 'control pixels', i.e. the answers have already been pre-determined by remote sensing experts. These pixels are taken to be the 'truth' and correctness is determined based on whether players agree

with the 'truth'. There are only a small number of control pixels in relation to the overall total pixels in the game.

The second way for determining correctness is through a 'majority rules' approach for those pixels where the answers are not known *a priori*. The first few times that a pixel is classified, the answer is always correct since a profile of answers must first be built up for each pixel. Correctness is then determined through agreement with the crowd at that point.

Figure 1: A screenshot from the Cropland Capture game.

The game will run for a total of 6 months from mid-November 2013 until May 2014 to provide a good compromise between collecting as much information as possible while still retaining participation. The incentives for participation are prizes awarded at the end of the game, which include smartphones and tablets. In order to be eligible for a prize, players must be included in the final draw for these prizes. To become part of the draw, players must be in the top three scores each week, where scores are reset to 0 on a weekly basis at midnight each Friday. Thus in total there will be 75 people in the final draw. Some individuals have made it into the top three more than once so they effectively increase their chances of winning the prizes at the end. Additional prizes will be offered during the last five weeks of game play to motivate additional participation.

The game was launched via a media campaign, with press releases, blogs and a twitter account set up for the game. From there, the game was picked up in a blog by National Public Radio, an article in the Guardian and reported with interviews by Geo-Wiki staff on German radio (Deutsche Welle) and Austrian media (ORF). All of these media outlets contacted us without any initiation by us. The Geo-Wiki network was also contacted, where we simultaneously launched a new monthly newsletter to provide regular updates on the game.

The game can be played online (see http://www.geo-wiki.org) or on an Apple or Android smartphone or tablet. The apps can be downloaded from the app stores for these devices.

Figure 2 shows the 'About Cropland Capture' screen, which provides some background to the game but also training materials. Players can view examples of cropland and land cover that is not cropland, which is useful for those players who have never viewed much satellite imagery before. There is also an FAQ, which contains answers to common questions that players have emailed us about.

Figure 2: The 'About Cropland Capture' screen.

Although not visible from Figure 1, the player's score will appear at the top and players can access a leaderboard to compare their progress against other players.

3 Initial Results

The location of the pixels is based on a global validation data set [18] at varying resolutions from 250m to 1km^2. We have also added in the locations at the Degrees of Confluence project (http://confluence.org). The dates associated with the imagery is recorded separately. The results presented here correspond to the data collected at around the halfway point of the game. At that point there were 2,817 players who contributed a total of 3,297,928 answers or image interpretations as part of playing the game. The same image was provided to many players as mentioned previously so there are currently 137,551 uniquely classified pixels of varying resolution. New pixels are always being added to the game.

Figure 3 shows the number of times that images have been classified multiple times. Roughly two-thirds of the images have been classified more than 5 times with one image having been classified more than 500 times.

Figure 3: The number of times that images have been classified multiple times

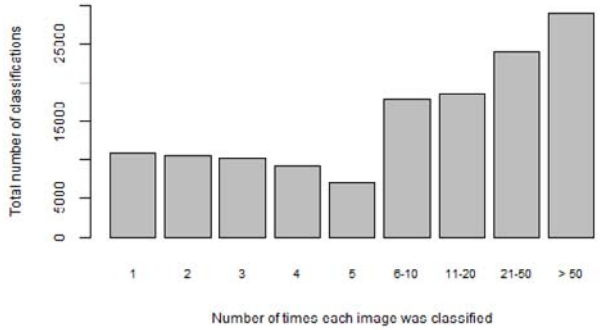


Figure 4 shows that the majority of images have a greater than 70% agreement between the players where only those images with more than 10 answers per image were included.

Figure 4: Example where agreement is less clear cut

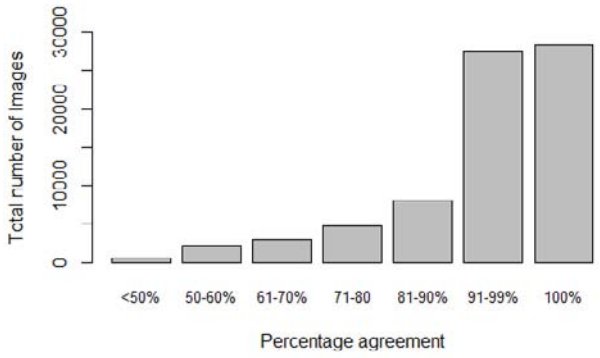


Figure 5 shows an example of high agreement on cropland. For this image 35 out of 36 players indicated cropland while 1 player said there was no evidence of cropland.

Figure 5: Example of cropland indicated 35 out of 36 times

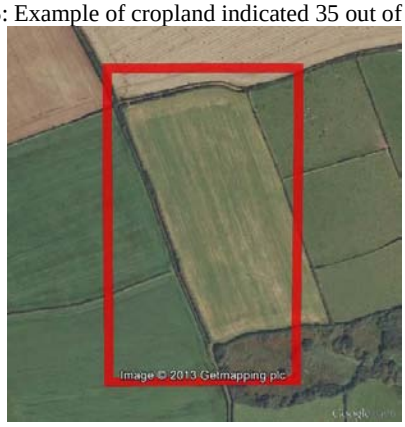


Figure 6, on the other hand, shows an example of an image where the players are split between cropland and non-cropland. Of the 59 evaluations for this image, 33 players said there was cropland, 25 said no cropland and 1 said maybe.

Figure 6: Example where agreement is less clear cut

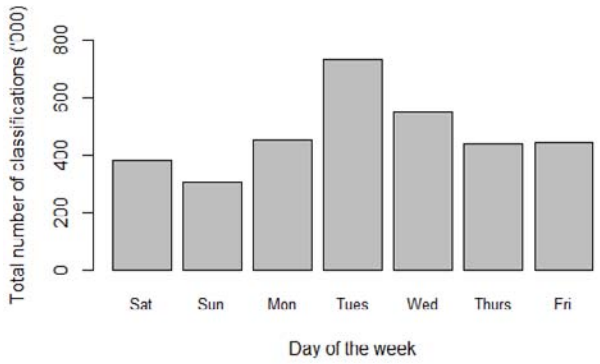
In this case the majority is still correct, i.e. there is cropland visible on the lower right hand side of the image.

Since the agreement with the crowd largely determines correctness (as outlined in section 2), a number of players contacted us about adding an option to disagree with the 'correct' answer provided. This option was added and is illustrated in Figure 7. For this image, 107 players said there was no evidence of cropland, 42 said cropland was present and 3 answered maybe. Using the majority rule, this image would contain no cropland as 'correct'. One of the players then disagreed with this answer and the image was automatically sent to an expert. The expert confirmed that cropland is present in the lower left corner of the image and the player who disagreed with the original answer of 'no cropland' was awarded extra points. This feature has now been used on various occasions to correct the answers provided by the crowd.

Figure 7: An example of an image contested by a player

There are many other analyses that are currently being undertaken with the data. A final example is provided regarding patterns of activity over the week. Since the competition scores are reset each week as explained in section 2, we expected the majority of activity to take place in the early part of the week, i.e. Saturday, Sunday and Monday as the players do battle for the top three positions early on during each week. However, as Figure 8 shows, this main activity occurs on Tuesdays and Wednesdays, which is counter to what we expected.

Figure 8: Number of classifications during the week



This type of temporal information could be used to provide a greater understanding of how incentives might affect player behavior over time.

4 Next Steps

We are currently in the process of analyzing the data for user performance, with the ultimate goal of developing simple rules that will determine the minimum number of classifications needed per pixel before we can be confident in the majority. Right now we have many pixels that have been classified more than 50 times yet it would be more efficient to remove pixels from the game when a minimum number has been reached, thereby allowing more areas to be classified. Right now the decision to remove pixels from the game is applied in an ad hoc basis but we are in the process of developing empirical rules for more efficient removal that are based on the results of the game so far. We will use these rules in future games.

Once the game is complete and the data filtered by quality, we plan to use the dataset for developing and validating a global hybrid cropland map where we will integrate many existing cropland products to produce a single, improved product. The data from the game will be used to help determine which product is correct at a given location and for validating the resulting map.

The success of the game so far means that we will use this type of approach for gathering other land cover types in the future.

Acknowledgements

EU COST Action TD1202 Mapping and the Citizen Sensor and the EU FP7 ERC Crowdland project.

References

[1] Albrecht, F., C. Perger, C. Schill et al. Using crowdsourcing to examine land acquisitions in Ethiopia. GI Forum, Salzburg, Austria, July 2, 2013.

[2] Bonney, R. et al. Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*. 59(11):977-984, 2009.

[3] Defourny, P., C. Vancustem, P. Bicheron et al. GLOBCOVER: A 300m global land cover product for 2005 using ENVISAT MERIS time series. *Proceedings of the ISPRS Commission VII Mid-Term Symposium: Remote Sensing: from Pixels to Processes*. Enscede NL, 2006.

[4] Deterding, S., M. Sicart, L. Nacke et al. Gamification. using game-design elements in non-gaming contexts. CHI 2011, Vancouver, BC, Canada, May 7-12 2011.

[5] dotMobi: Global mobile statistics. Section E: Mobile apps, app stores, pricing and failure rates, http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats/e#popularappcatagories, 2013.

[6] Friedl, M.A., D. Sulla-Menashe, B. Tan et al. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets, *Remote Sensing of Environment*, 114(1): 168-182, 2010.

[7] Fritz, S. et al.: Downgrading recent estimates of land available for biofuel production, *Environ. Sci. Technol.*, 47(3):1688-1694, 2013.

[8] Fritz, S., I. McCallum, C. Schill et al.: Geo-Wiki.Org: The use of crowdsourcing to improve global land cover, *Remote Sensing*. 1(3):345-354, 2009.

[9] Fritz, S., I. McCallum, C. Schill et al.: Geo-Wiki: An online platform for improving global land cover, *Environmental Modelling & Software*, 31:110-123, 2012.

[10] Fritz, S., E. Bartholomé, A. Belward et al. *Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version)*. Office for Official Publications of the European Communities, Luxembourg, 2003.

[11] Goodchild, M.F. Citizens as sensors: the world of volunteered geography. *GeoJournal*. 69(4):211-221, 2007.

[12] Howe, J. The rise of crowdsourcing. *Wired Magazine*, 2006.

[13] Michael, D.R. and S.L. Chen. *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade, 2005.

[14] Pittman, K., M.C. Hansen, I. Becker-Reshef et al. Estimating global cropland extent with multi-year MODIS data, *Remote Sensing*. 2(7):1844-1863, 2010.

[15] See, L., S. Fritz, C. Perger, et al. Mapping human impact using crowdsourcing. In: Carver, S. and Fritz, S. (eds.) *Mapping Wilderness: Concepts, Techniques and Applications of GIS*. In press Springer, 2014.

[16] Sturn, T., D. Pangerl, L. See et al. Landspotting: A serious iPad game for improving global land cover. GI-Forum, Salzburg, Austria, July 2 2013.

[17] Yu, L., J. Wang, N. Clinton et al. FROM-GC: 30 m global cropland extent derived through multi-source data integration, *International Journal of Digital Earth*, 6(6):521-533, 2013.

[18] Zhao, Y. et al. Towards a common validation sample set for global land cover mapping. *Photogrammetric Engineering and Remote Sensing*. In review.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Applying a CA-based model to explore land-use policy scenarios to contain sprawl in Thessaloniki, Greece

Apostolos Lagarias
Aristotle University of Thessaloniki (AUTH) & Regional Analysis Division, Institute of Applied and Computation Mathematics, Foundation for Research and Technology - Hellas, 100 N. Plastira, Vassilika Vouton, 70013 Heraklion, Crete, Greece
lagarias@iacm.forth.gr

Poulicos Prastacos
Regional Analysis Division, Institute of Applied and Computation Mathematics, Foundation for Research and Technology - Hellas, 100 N. Plastira, Vassilika Vouton, 70013 Heraklion, Crete, Greece
poulicos@iacm.forth.gr

Abstract

This study addresses the issue of urban sprawl through the application of a Cellular Automata (CA) based model in the area of Thessaloniki, Greece. To link macro-scale to micro-dynamic processes the model integrates a statistical model at the regional level with a CA model at the local level. The model is used to compare two scenarios of growth of Thessaloniki to year 2030; the first one assuming a continuation of existing trends, whereas the second one assuming the enactment of various land use regulations in order to contain urban sprawl. The comparison of the results demonstrate that in the second scenario there is a smaller degree of leapfrog growth, with high percentage of new developed land being inside the existing city plans with development in areas outside the plans and in agricultural areas being minimized.

Keywords: Cellular Automata, urban sprawl, urban modelling, land-use policy, Thessaloniki

1 Introduction

Urban sprawl represents urban growth characterized by a sharp imbalance between urban spatial expansion and the underlying population growth *[1]*, discontinuous spatial development patterns *[5]* and low densities *[3]*. Critics of sprawl have emphasized its negative impacts and especially the fact that it leads to increasing car-dependency for transportation, need for more infrastructure, loss of agricultural and natural land, higher energy consumption and degradation of periurban ecosystems *[11]*.

This paper addresses the issue of urban sprawl through the application of an urban growth model based on Cellular Automata (CA). Spatially explicit urban expansion models can effectively trace development patterns of the past and assess possible expansion/future scenarios, they can be therefore regarded as important tools in urban sprawl analysis and could be of use to planners *[13]*. The model takes into account a wide range of demographic, accessibility, socioeconomic, environmental and urban planning data, as well as, a set of local characteristics at the cell level.

The model's logic, structure and calibration procedure are briefly outlined in this paper. A more detailed discussion is provided in *[8]*. The case study area is the urban agglomeration of Thessaloniki in Greece, an area in which the traditional monocentric and radial growth has been supplanted in the last 20 years by rapid periurbanization and extensive urban sprawl. The model is used to compare two growth scenarios for the period 2010-2030; one assuming a continuation of the existing trends and the second one assuming the adoption of various land use regulations with the objective to contain urban sprawl.

2 Methodology

2.1 CA models and urban sprawl

Cellular automata (CA) is a class of spatially disaggregate models consisting of a two-dimensional lattice of cells, in which each cell is characterized by a particular state determined by a set of transition rules *[14]*. Each cell corresponds to a patch of land, and the state(s) of the cells represent the different land-uses. CA are discrete, iterative and dynamic mathematical constructs in which the state of each cell depends on its previous state and on the state of the cells in its neighbourhood *[15]*.

Urban sprawl is a dynamic phenomenon that can be modelled through an analysis of land use and land cover changes *[10]*, therefore extensive research has been carried out to demonstrate the CA's model suitability for simulating urban expansion and sprawl. Among various widely applied models reference can be made to SLEUTH *[2, 7]*, MOLAND *[9]*, DUEM *[18]*, the SimLand model *[17]*, the model developed by He *et al. [6]*, the model developed by Torrens [12] etc.

2.2 Model framework and structure

In the present context a CA-based model is applied in the region of Thessaloniki, to explore future development scenarios. The model integrates a statistical model at the regional level with a CA model at the local level, in order to

capture macro-scale and micro-dynamic processes. The model therefore presents a two level framework for simulating urban growth, taking into account macro-scale characteristics at the aggregate zone level (municipality or district level) and place-specific, local characteristics at the cell level. The statistical model takes the form of a linear regression function of the type

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots + b_nX_n \quad (1)$$

where Y represents the new urban (developed) land located in zone k, and X_1, X_2,…X_n various explanatory factors (population change, land availability, accessibility, socioeconomic-demographic data, land values, environmental information etc.) and $b_1, b_2, \ldots b_n$ the estimated coefficients.

At the cell level for estimating propensity of development two indexes are used. The first one is similar for all cells within a zone and represents the urbanization potential of the zone. It is defined as:

$$P_{GLk} = ULC_{ch(k)} / \sum_{k=1}^{m} ULC_{ch(k)} \quad (2)$$

where $ULC_{ch(k)}$ urban land cover change in each zone k as estimated by equation (1) and m the total number of zones. The second one, represents the cell urbanization potential based on its location (slope), land use regulations (protected areas), planning regulations (area of buildings allowed to be constructed), adjacency to other developed areas, proximity to major transportation corridors and proximity to major settlement centers. Mathematically, it is expressed as:

$$P_{LCi} = PR_i \cdot G_i \cdot Z_i \cdot (ND_i + RD_i + D_i) \quad (3)$$

where P_{LCi}=the (local) urbanization potential of a non-urban cell i; PR_i = a protection (construction permission) factor according to land use regulations that takes the value of 1 for areas where development/construction is permitted and the value of 0 for protected areas in which development is prohibited; G_i=a geomorphology factor that is equal to 1 in areas where land slope permits building and 0 in steep slope mountainous areas where construction is difficult; Z_i = a cell specific parameter that expresses the "building ratio/factor" that is the max allowed area of the building (in all floors) as defined by the planning regulations; ND_i = a factor describing the density of surrounding urban uses, related to sprawl-coefficient and to a compact-coefficient index, RD_i = a road network proximity factor related to a road-coefficient index; D_i = a distance factor from the nearest settlement center, related to a distance-coefficient index. The $P_{LCi\ I}$ index is normalized so that $\Sigma P_{LCi(k)} = 1$ for i =1, 2… N where N is the number of cells (i) in zone k.

The overall cell urbanization potential for a cell *i* in zone *k* $P_{i(k)}$, results from the multiplication of the two indexes expressed as:

$$P_{i(k)} = P_{GL(k)} \cdot P_{LCi(k)} \quad (4)$$

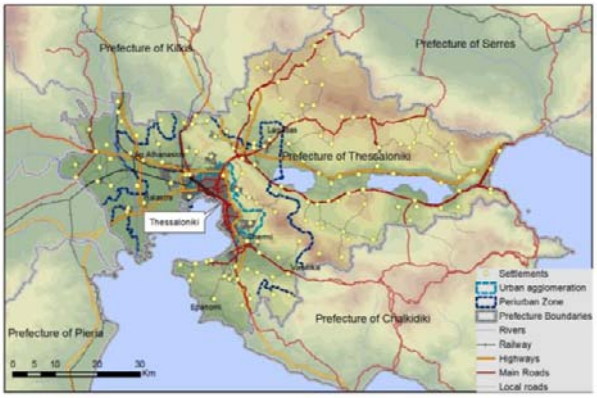
The model has been developed in the Netlogo *[16]* cellular automata and multi-agent programmable modelling environment. The model is loosely coupled to GIS through its inputs and outputs. Data on geomorphology, protected areas, land use zoning status etc. are stored in the GIS and imported to the model through ASCII files. The results of the simulation are exported as a JPEG file and imported in the GIS database for further analysis as a georeferenced raster image.

2.3 Data and calibration procedure

Thessaloniki is the second largest metropolitan area in Greece. It is the capital of the region of Central Macedonia with a population of almost one million people. In the study area (urban agglomeration and periurban zone) there are 29 municipalities, with the city of Thessaloniki being the largest one, and 56 different city-districts. The area, traditionally characterized by a densely built-up and monocentric structure, has seen since the '80s a change in the growth patterns with development in the periurban areas, with population decentralization and rapid urban land expansion. Urban land (developed area) increased between 1990 and 2010 by nearly 5.000 ha, growing by approximately 60% at an average rate of 2.5% per year, while at the same time population increased by 160.000 people, that is almost 20%.

Land cover data were obtained from satellite images for years 1990, 2000 and 2010 (Spot and Landsat images at a resolution 20-30 meters). Essential ancillary data, including 1:50.000 topographic maps, boundaries of protected areas and planning regulations with respect the "building/ratio" factors were obtained from local planning offices. The major transportation network, contour lines, land use zones and other necessary data were digitized from paper maps from various sources. All data layers were registered to the Greek coordinate system (EGSA87). Statistical data were obtained from the National Statistical Service of Greece from the Censuses of 1991, 2001 and 2011.

Figure 1: Case study area within the region of Thessaloniki



The multiple regression model (eq. (1)) estimating developed land change was calibrated for the 1991-2011 time period at the district level. As shown in Table 1 urban expansion is positively related to land available for urbanization (AVarea), population growth (POPrat), population density change (PDENch), population in the age cohort 15–64 years (Fdem) and to areas with significant concentration of services oriented establishments (SERVrat), while it is negatively related to the distance from the city centre (InDis) (city centre was defined to be in Aristotle's square, in Thessaloniki). The variables were inserted in the model through a stepwise process. R square coefficient is estimated to be 0.79, with all variables being significant at the 5% statistical level.

Table 1: Multiple regression results

Variable	Estimate	St.E.
AVarea	0.762***	0.108
POPrat	0.533***	0.157
PDENch	0.002**	0.001
Fdem	0.009**	0.005

SERVrat	0.201***	0.068
InDis	-0.039***	0.010
Constant	0.074**	0.031

**indicates significant at a=0.05 ***indicates significant at a = 0.01

For the calibration of the CA model the area was subdivided into 100x100 meter cells each cell therefore representing an area of 1 ha. A total of 120,536 cells were defined and a 3-cell radius neighbourhood was used in the model. The model was calibrated by running simulations for the period 1990-2010. The calibration process involves running the model forward and comparing the resulting map to the observed land cover data, until converging to a set of coefficient scores that match a set of goodness-of-fit criteria that measure the similarity of the observed and simulated patterns. Five different indexes/metrics were used to test goodness-of-fit of the model, namely a) Index of Compactness, b) Index of Sprawl, c) Index of Road-driven Development, d) Index of Incorporation, and e) Index of Concentration. These indexes express respectively the proportion of new urban land located in areas with high local density of urban land uses, in areas with low local density, in areas adjacent to the road network, in areas incorporated in the city plan and in areas within a two-kilometer radius from the nearest settlement center.

According to the calibration procedure, the sprawl coefficient value is relatively high (sprawl-coef=0.80) while the compact coefficient takes a smaller value (compact-coef=0.20). The significance of planning regulations and plot ratios for attracting development is reported through the building ratio coefficient score (zone-coef=0.20). Important is also the road-driven development (road-coef=0.15) and the role of existing settlement centres (distance-coef=0.10).

The model fits well to the observed changes for the period 1990-2010, as all five of the goodness of fit indexes deviate less than 5% from the observed values. The Absolute Matching Index is equal to 0.38, stating that 38% of observed new urban cells are predicted by the model at their exact location, while the Matching Index with Tolerance is equal to 0.91, stating that 91% of observed new urban cells are predicted by the model at a location closer than 300 meters (same as the size of the CA neighbourhood) from their exact location.

3 Scenarios of sprawl

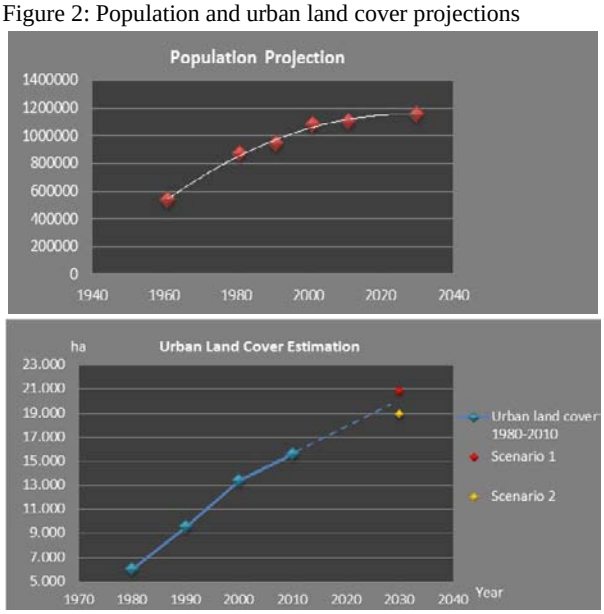
3.1 Scenario assumptions

The calibrated model is used to compare two scenarios for the period 2010-2030, the first assuming continuation of the existing trends - and the current land use planning and policy context, and the second a more sustainable development scenario based on land use zoning policies aiming to contain sprawl.

For both scenarios total population increase and the corresponding urban land cover growth were first estimated and then the model was used to allocate urban growth to the individual cells. To estimate total population increase and age distribution for year 2030 several assumptions were made extrapolating past patterns and using national averages. Most importantly the immigration rates were assumed to be significantly lower than those in the past. Overall, it was estimated that population will increase by about 50,000 people that is about 5%.

To estimate the corresponding urban land growth the concept of the Sprawl Index was used. As defined by EEA *[4]* the Sprawl Index is the ratio of the rate of urban land cover change divided by the rate of population change. Values higher than 1 indicate a sprawling process while values lower than 1 are related to compact development. In the study area the sprawl index during the 1990-2010 period was equal to approximately 3.0 while in the period 2000-2010 it reached a value of almost 10. On the basis of these growth patterns and taking into account a moderate growth scenario as a result of the recent economic recession in Greece, it was assumed that the Sprawl index in the first scenario will not exceed the value of 6 and in the second scenario the value of 4. As shown in Fig 2 both population and land cover estimations follow recent growth trends. Converting these growth assumptions into developed land acreage it is estimated that the land to be urbanized/developed in the two scenarios is 5,255 ha and 3,340 ha respectively.

Figure 2: Population and urban land cover projections



3.2 Scenario 1: Continuation of existing trends (Bussiness-as-usual)

In scenario-1 the following assumptions were made: a) continuation of existing land use/planning regulations with respect construction/development outside the official settlements' boundaries in non-protected areas b) expansion of the city plan permitting the construction of up to 9,500 ha of residential, commercial-housing buildings, c) population decentralization with population growing by up to 2.5% per year in selected periurban settlements and decreasing in the central city zone by 1% per year d) continuation of the existing trends of the location of commercial, office and entertainment facilities in periurban areas and mainly in the south-eastern sector of the city e) Construction of new road infrastructure, as anticipated by the new regional plan of Thessaloniki and the relevant traffic and transportation studies.

According to existing planning regulations, new residential areas are expected to develop as low density with small plot ratios. Low density development is also favored by the land

use regulations which permit development on agricultural areas, as long as the lot has a minimum area of 4 ha therefore practically encouraging uncoordinated urban land expansion in agricultural land.
For the allocation of the 2030 developed land to the cells the calibrated equations were used. Results (table 2) show that urban sprawl will continue at rapid rates, with existing designated areas attracting only a small percentage of new urban land (Incorporation Index equal to 36%), while 23% being located in low density neighborhoods (Sprawl Index equal to 23%).

3.3 Scenario 2: Land use policy restrictions to contain sprawl

In scenario-2 the impacts of a land use policy aiming to contain sprawl and to promote a more compact urban growth were examined. It was assumed that: a) there would be smaller expansions of the city plan for residential use, with higher plot ratios than in scenario-1, b) restrictions with respect building construction outside the plan areas will be imposed, c) new protection areas will be established mostly on high yielding agricultural land, mainly in southern periurban zone and in the west periurban zone areas, as well as on the northern mountainous areas. Population decentralization will continue at lower rates and "central" uses will be concentrated in selected large periurban settlements.
Simulation results (table 2) show a smaller degree of urban sprawl, with 45% of new urban land being incorporated into existing designated areas, 19% of new urban areas being located in low density neighborhoods, 29% adjacent to the major road network and 66% located within a distance of 2 km from the settlements located in the periurban zone.

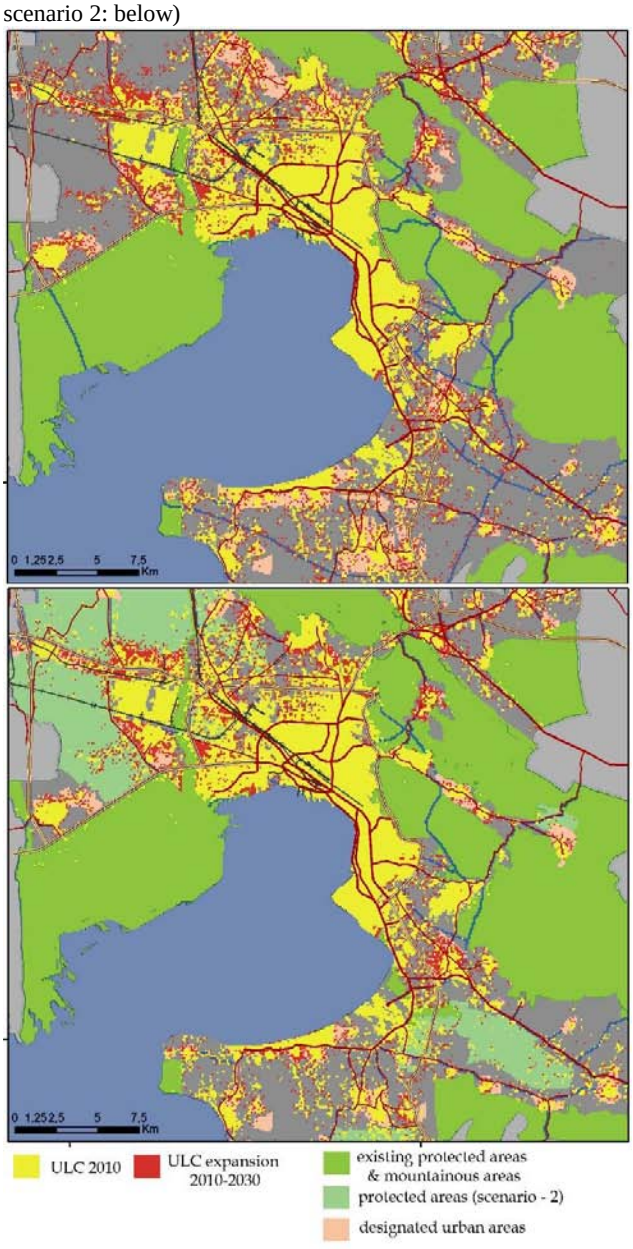
Table 2: Growth trends for both scenarios

Index	Scenario-1	Scenario-2
Compactness	77%	81%
Sprawl	23%	19%
Road-driven	31%	29%
Incorporation	36%	45%
Concentration	64%	66%

3.4 Results and discussion

In this study a CA-based model of urban growth was used to compare two future development scenarios in the metropolitan area of Thessaloniki, one assuming continuation of existing trends and another one aiming to contain urban sprawl through the adoption of land use regulations. The growth patterns obtained differ to some extent. In scenario-2 there is a lesser degree of leapfrog development with a high percentage of new urban areas occurring inside existing boundaries, while urban sprawl in environmentally important areas and fertile agricultural land is reduced.
Overall it can be stated that spatially explicit urban models can successfully trace development patterns of the past and can be used to assess growth patterns of the future. If appropriately articulated to include variables that are affected by land use policies drawn by decision makers, then they can be used not only for simulating the trends, but most importantly for simulating the impact of alternative policies. They are, therefore, valuable tools for analyzing urban growth and urban sprawl and further development of CA models should concentrate on enriching these models with variables and relationships that can be used for testing the impact of alternative policy measures with respect future urban growth.

Figure 4: Scenario simulation results (scenario-1: above / scenario 2: below)



Acknowledgments:
Parts of the research reported in this paper has been funded by the PEFYKA project of the General Secretariat for Research and Technology.

References

[1] J. Bruekner. Urban Sprawl: Lessons from Urban Economi*cs. Brookings-Wharton Papers on Urban Affairs*, 2001

[2] K.C. Clarke L. Gaydos S. Hoppen. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B* 24: 247 – 261, 1997

[3] C. Couch., L. Leontidou, G. Petschel-Held, editors. *Urban sprawl in Europe: Landscapes, Land-use change & policy*. Wiley-Blackwell Publishing, 2007

[4] Environmental European Agency (EEA). *Urban sprawl in Europe: The ignored challenge*. EEA Report No 10, 2006

[5] R. Ewing, R. Pendall, D. Chen. *Measuring Sprawl and Its Impact*, Smart Growth America, 2002

[6] C. He, N. Okada, Q. Zhang, P. Shi, J. Zhang. Modeling urban expansion scenarios by coupling cellular automata model and system dynamic model in Beijing, China. *Applied Geography* 26: 323–345, 2006

[7] C. Jantz, S. Goetz, Shelley M. Using the SLEUTH urban growth model to simulate the impacts of future policy scenarios on urban land use in the Baltimore-Washington metropolitan area. *Environment and Planning B: Planning and Design* 30: 251 – 271, 2003

[8] A. Lagarias. Urban sprawl simulation linking macro-scale processes to micro-dynamics through cellular automata, an application in Thessaloniki, Greece. *Applied Geography*, 34: 146-160, 2012

[9] C. Lavalle, J. Barredo, N. McCormick, G. Engelen, R. White, I. Uljee. *The MOLAND model for urban and regional growth forecast. A tool for the definition of sustainable development paths*. European Communities, Joint Research Centre, 2004

[10] A. Schneider and C. Woodcock. Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information. *Urban Studies* 45: 659-692, 2008

[11] P. Torrens and M. Alberti. Measuring sprawl. *CASA working paper series 27*, UCL, 2000 http://www.casa.ucl.ac.uk/publications/workingpapers.asp , last accessed 1/1/2013

[12] P. Torrens. Simulating Sprawl. *Annals of the Association of American Geographers* 96 (2): 248–275, 2006

[13] D. Triantakonstantis, P. Prastacos, A. Tsoukala. Analyzing Urban Sprawl in Rethymno, Greece. *J Indian Soc Remote Sens*, 2014

[14] R. White, G. Engelen. Cellular automata and fractal urban form: a cellular modeling approach to the evolution of urban land-use patterns. *Environment and Planning A* 25: 1175-1199, 1993

[15] R. White, G. Engelen, I. Uljee. The use of constrained cellular automata for high-resolution modeling of urban land-use dynamics. *Environment and Planning B* 24: 323-343, 1997

[16] U. Wilensky. NetLogo. http://ccl.northwestern.edu/netlogo/. Centre for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL, 1999

[17] F. Wu Simland: a prototype to simulate land conversion through the integrated GIS and CA with AGP-derived transition rules. *International Journal of Geographical Information Science*, 12: 63-82, 1998

[18] Y. Xie and M. Batty. Integrated urban evolutionary modeling. *Casa working papers, Paper 68*, 2003, http://www.casa.ucl.ac.uk/publications/workingpapers.asp last accessed 1/1/2013

Session:
Trajectories

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Queues in Ski Resort Graphs: the Ski-Optim Model

Tino Barras
University of applied Sciences western Switzerland
Rue du technopôle 3;
3960 Sierre
Sierre, Switzerland

Marut Doctor
University of applied Sciences western Switzerland
Rue du technopôle 3;
3960 Sierre
Sierre, Switzerland
Marut.doctor@hevs.ch

Marc Revilloud
University of applied Sciences western Switzerland
Rue du technopôle 3; 3960 Sierre
Sierre, Switzerland

Michael Schumacher
University of applied Sciences western Switzerland
Rue du technopôle 3;
3960 Sierre
Sierre, Switzerland
Michael.schumacher@hevs.ch

Jean-Christophe Loubier
University of applied Sciences western Switzerland
Rue du technopôle 3;
3960 Sierre
Sierre, Switzerland
jchristophe.loubier@hevs.ch

Abstract

It is rather unknown how skiers move inside ski areas. However, new data collection systems, such as RFID chips on ski passes (which allow counting skiers at the gates of the cableways), can be used to analyse the movement of skiers in the cableways network and in the ski runs graph. This will show how queues arise at the cableways departures and how crowds are formed on the ski runs. This short paper is reporting a multi-agent simulation approach called Ski-Optim to study graphs and queues arising in a ski area. A software simulation was experimented on the ski area of Verbier in Switzerland.

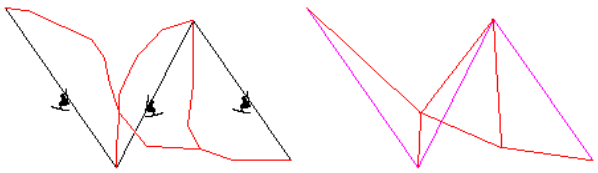
Keywords: Multi-agent simulations ; flows; queues ; ski area ; graphs

1 Introduction

The Ski-Optim model was elaborated for the Swiss ski area of Verbier. This place was selected for this project as it is entirely provided with the Ski Data system (management system RFID of the gates at the cableways) and also with a georeferenced plan of the ski runs. Besides, Verbier has the advantage to be one of the biggest ski resorts in Switzerland with a big number of skier days (950'000 in 2013), a complex network of 34 different cableways and 195 km of ski runs. Thus, the model can be tested in a sufficiently big environment.

Every ski resort can be modelled from a unique basic form with one ski lift, one toggle and one ski run [1]. Formally, a ski resort is a connected oriented graph. The axis properties are specific and influence the behavior of the agents (the skiers), which will follow a way as a function of their projet of ski (Figure 1).

Figure 1 : Flow graph of a ski area.



Crossings at ski lifts is a typical queue problem. Skiers are clients who arrive randomly in a waiting zone composed of one or several ski lift (called servers in queue theory [2]). The service duration (the skier on the ski lift) is also random and when all the servers are occupied, a queue appears with a waiting, which corresponds (the most of the time) to first come, first served [2]. Now, the question is how to handle these phenomena in a network such as a ski area because this problem is important due to potential influences of negative skiing experiences.

2 Mechanism of the Running of a Ski Area: Characterization of the Skiers Behavior

To simulate skiers' behaviors in a ski resort, it is necessary to model the load increase of skiers in the network and to model the skiers' diffusion on the ski runs [3].

In a ski resort, the schedule of a day is uniform. In a first step, skiers arrive and cross the snow front (entrance point in ski areas). Then, they go up with the cableways until the network point that corresponds to the departure of the specific ski run they wish to go to.

Skiers have action plans (which will be used in the simulation). They can decide for example to reach the highest point of the ski resort or choose a precise sector according to their plan. As the access to the ski runs depends on the cableways, the load increase in the network is not

homogenous. There is a time lag for the cableways and ski runs which are reachable by only other cableways or other ski runs. This leads to jam processes at the beginning of the day. After some time, following a dilution process of all skiers in the network of the ski resort, a steady state arises. This can explain why the cableways with the biggest transport capacities are located at the beginning of the network where the ski runs are also broader. These infrastructures should be able to absorb a punctually great number of skiers at the beginning and at the end of the day. Due to this collective effect, they are also the most used ski runs in the course of the day.

As most of the ski resorts are provided with RFID systems for getting through cableways turnstiles, it is nowadays possible to obtain the number of skiers on the cableways according to any time slot. It is however still impossible to measure the number of skiers on the ski runs. The local behavior (ski run selection and path) of the skier is therefore unknown. So, the information put on the graph is incomplete. It leads to difficulties for the processing of the queues by an analytic method in classical queue theory. As we cannot use an analytic method, we use a multi-agent-based simulation.

3 Approach by Multi-Agent-based Simulation: Ski-Optim

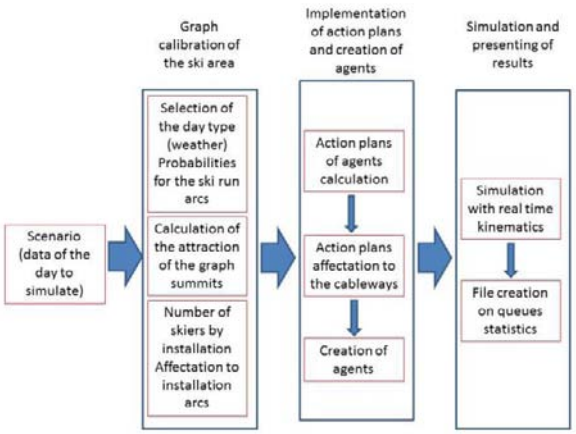
A multi-agent-based system can be defined by a set of processes running simultaneously, exchanging information mutually and sharing common resources [4]. These processes are called "agents". In our case, there are several agents using at the same time the ski runs. Those agents do not interact directly with each other.

3.1 General Principles of Ski-Optim

In the Ski-Optim model, an agent follows an action plan [4], which it gets at the departure and which it will not modify during his whole day on the ski runs. The simulation corresponds thus to a daily pattern. The daily simulation can move a great number of agents on the ski runs graphs. Furthermore, the model relies on the counting of data at the turnstiles of the cableways. This information is useful to simulate the influence of the flow of skiers depending, on one hand, on the flow rate of every cableway and, on the other hand, on the location of the cableways in the graph. We can therefore follow the change in waiting time at the departure of every cableway on the basis of the number of skiers using the cableways during the day. It is also possible to follow the change of the skiers flow for each ski run segment. This permits to proceed for adjustments of flow rates and speed at the different cableways and then to measure the impacts during the simulation.

The figure below shows the complete simulation process. The following sections explain those steps.

Figure 2 : Mechanism of the simulation progress.



3.2 Graph Calibration

The calibration process of the graph is one of the most important part of the simulation as it is at this level that the diffusion rules in the network are implemented. The diffusion factors are only partly known in the network (only on the cableways) as it is impossible to give a measurement of the number of skiers on the ski runs. This is why a probability selection on the ski runs used on the form of rules is inserted (Figure 3). These rules are defined according to the weather (simplified by four types of weather situation: sunny/warm, sunny/cold, rain, snow). They were created as a result of a survey on skiers' behavior.

Figure 3 : Definition rules on probabilities of ski run use.

```
if(jourType == 1) // Sunny and warm
{
   blueProba = 0.27f;
   redProba = 0.42f;
   blackProba = 0.23f;
   yellowProba = 0.08f;
}else if(jourType == 2) // Sunny and cold
{
   blueProba = 0.09f;
   redProba = 0.36f;
   blackProba = 0.25f;
   yellowProba = 0.25f;
}else if(jourType == 3) // Rain
{
   blueProba = 0.27f;
   redProba = 0.44f;
   blackProba = 0.21f;
   yellowProba = 0.08f;
}else if(jourType == 4) // Snow
{
   blueProba = 0.23f;
   redProba = 0.46f;
   blackProba = 0.21f;
   yellowProba = 0.10f;
}
```

This operation is performed manually via an icon set which gives the opportunity to decide in which weather frame the simulation will take place. The result of the operator selection is the assignment of the weightings on the ski runs arcs of the ski area graph. This is followed by the calculation of the attraction coefficient from the summits to the departure of the cableways. For this purpose, the model uses an algorithm (described in [3]) which calculates the attraction coefficient from each departure point from the summit of a cableway, depending on the number of skiers which used that cableway according to the counting done at the turnstiles. This coefficient is the indicator which gives the opportunity to determine if a departure of a cableway is more used than another in the course of an hourly time slot. This coefficient is recalled for every temporal iteration. This algorithm calculates also the action plan of an agent, depending on the available ski run type combined with the selection of the day type by the operator.

3.3 Implementation of Action Plans and Creation of Agents

The model agents are simplified and do not interact independently. They follow an action plan, which determines the way to go from a point A to a point B.

Before calculating these action plans, three steps are necessary:

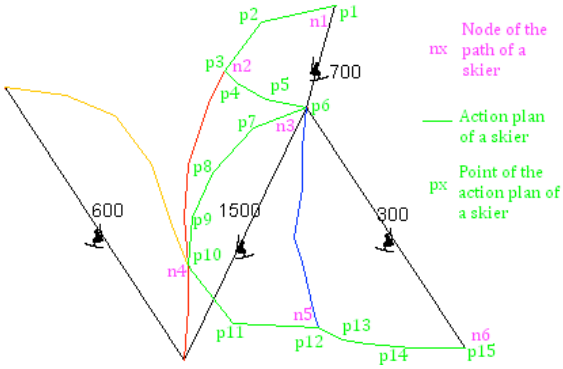
1. Collection of all the counting file data at the cableways turnstiles;
2. Calculation of the difference of the number of agents in comparison to the preceding period, in order to keep the total number of agents on the graph;
3. Values update depending on the principle cableways (in the particular case : entry and exit points)

Secondly, the action plans are created, grouped by departure point, mixed and inserted in a list. Each agent created on the summit of a cableway gets an action plan and unrolls it until his point of arrival (figures 4 and 5).

Figure 4 : Pseudocode of the action plans management

```
FOR EACH action plan corresponding to one cableway (APs)
    FOR EACH cableway
        IF the departure of the cableway = first point of the first of the APs THEN
            Add the APs in the list of the cableway
        END IF
        IF the saving of the action plans < the current list THEN
            Save the list of the action plans for
            1. The last cableways
            2. When there are no more action plans available anymore
        END IF
        IF The cableway has action plans THEN
            Shuffle the list of the action plans
        END IF
    END FOR EACH
END FOR EACH
```

Figure 5 : Example of an action plan of a skier in a graph



3.4 The Case of Entry and Exit Points

The number of skiers is known at the turnstiles of the cableways. However, a double counting should be eliminated, in order to add or withdraw only skiers who actually enter or come out of the ski area in the next time slot. For this purpose, the counting done at the main cableways are used, i.e. in the cableways where skiers are obliged to pass by for coming in or going out of the ski area. A cableway is considered "main" when the number of crossing is greater than other cableways. It is also assumed that more agents are going to the bottom of this main cableway. The calculation depends if the agents enter or go out from the ski area but the general mechanism is identical and is based on the difference of counting with the previous hourly time slot. This result is used to correct the number of RFID counting of a main cableway, so that the result remains coherent. In the case where there are several main cableways in the ski area, the model splits the new skiers and the skiers going out of the ski area proportionally to these main cableways according to the raw RFID counting.

(1) the case of the incoming skiers

Generally, skiers entering choose ski lifts at the interface with the snow front. The input mechanism is measured by time slots. At T-1, it is possible to calculate the frequency of usage based on volumes measured by the RFID chips. These frequencies allow, by a simple multiplication of the volume by frequency, to distribute correctly on each main ski lift; new skiers enter at T +1.

In addition, new agents are created at the departure of cableways at regular intervals in the ongoing hourly time slot.

These new numbers of agents will be used for the calculation of the action plans. In addition, new agents are created at the departure of cableways at regular intervals in the ongoing hourly time slot.

(2) The case of outgoing skiers

When the number of skiers decreases, as for example at the end of the day, the counting should also be adapted. As at this time, the skiers using the main cableways are less numerous than the skiers at the exit point, the outgoing skiers have to be added to the RFID counting according to an equivalent but inverse process, which has been described for the incoming skiers. In a second step, the number of agents to delete is assigned to the cableways in question. Thus, the model deletes

the foreseen agents at the different departures of the main cableways at regular intervals during the following hourly time slot, by distributing uniformly the number to delete. This creates a decimal number of agents, depending on the time slot. As a number of agents can only be an integer, the decimal part is stored in a variable created specifically (agentsFloat), in order to conserve the information. The calculation is done according to the following pseudo-code (figure 6).

Figure 6 : Pseudocode of management of the remains in the deleting of the outgoing agents

```
FOR EACH Cableway
  IF the cableway has agents to delete THEN
    IF the cableway has no agentsFloat to delete THEN
      agentsFloat = number of agents to delete / number
    of steps for 1h simulated
    END IF
    The value is stored after the coma of agentsFloat
    An integer is withdraw if >= 1.00
    This integer is added to the number of agents to
    delete during this step
  END IF
END FOR EACH
```

Furthermore, every time that an agent arrives at the departure of a main cableway, the model should test on one hand, if the cableway is closed, and on the other hand, if the agents have to leave the ski area at this point or if the agents have to be inserted in the queue (figure 7).

Figure 7 : Pseudocode of the selection between the deleting (exit of the ski area) or the insertion in the queue

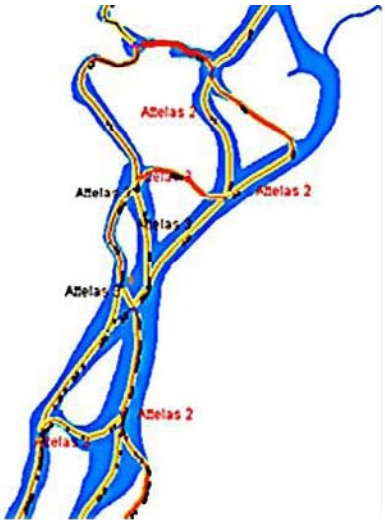
```
FOR EACH cableway
  IF the agent is located at the bottom of this cableway
  THEN
    IF the cableway is closed
        AND agents to delete THEN
      The agent leaves the station
      The number of agents to delete of that
      cableway is decremented
    END IF
    IF remains agents to delete
        AND remains agents to delete for this
      time slot THEN
      The agent leaves the ski area
      The number of agents to delete at that
      cableway is decremented
      The number of agents to delete during this
      slot is decremented
    END IF
    ELSE
      IF the skier doesn't exist yet in the queue
        The agent joins the queue
      END IF
    END ELSE
  END IF
END FOR EACH
```

This approach lets appear naturally the queue phenomenon by a simple test process. It corresponds to the counting of agents at the departure point divided by the hourly rate of flow of the cableway. The time variation in the waiting time is presented in real time in the kinematics simulation.

3.5 Simulation and Results

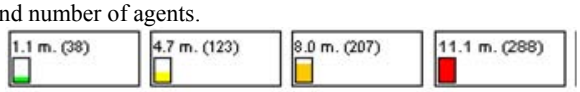
The Ski-Optim model simulates the congestion of the ski area with an hourly granularity as well for the queues as for the ski runs arcs. Figure 8 shows on the ski runs the emergence of zones congested by skiers (by the gradient variation from the green to the red)

Figure 8: Example of simulation results of the ski runs. The black points are the agents in action.



The simulation kinematics shows the change of the waiting time in the queues in real time by using the pictograms put on all the departure points of the cableways (figure 9).

Figure 9: Pictograms showing the waiting time in minutes und number of agents.



The values above the pictograms show the waiting time (without the brackets) in minutes and the number of agents in the queue (with the brackets). A statistics file is generated at the end of the simulation. It lets transcribing the change with time of the queues for every cableway of the graph for every time slot. In that way, the ski resort managers can perform the analysis on the potential global impacts (at the scale of skier diffusion in a graph) of a local choice of a graph property change (for example the change of an hourly rate of flow of a cableway). Figure 10 shows this output file of the simulation.

Figure 10: Output file generated from the simulation (zoom on hour no 5)

	Max waiting time	Min. waiting time	Average waiting time	Max agents on queue
Hour5				
Attelas1	.		.	1
Attelas2	3.2		1.6	109
Chaux2	.1		.1	4
Chxexpch				
Chxexpru	.7		.4	13
Combe1	.6		.3	11
Jumbo	.6		.3	10
LacVaux1	.4		.2	12
LacVaux3	.3		.1	5
Mayentzet	.9		.4	12
Medran	.8		.4	25
Ruinettes	13.5	11.8	12.6	344
Hour 5 averages	1.8	1.	1.4	

This file is an overview of the queues behavior at the cableways. It can be known for every hourly time slot, the

longest, shortest and mean waiting time for the analyzed hour and the maximum number of agents in the queue. This information is of course available for any hourly time slot of the day and for any cableway.

Discussion: Ski-Optim model provides some answers on the mechanisms like the emergence of the phenomenon of queuing and packet skiers on the slopes. At the operational level, we observe that the diffusion in the graph is strongly impacted in secondary levels (accessible only lifts from other lifts) when the main lifts have their carrying capacity increased.Work will be undertaken to try to determine if this effect follows a probability law and if it is disturbed by the change in the structure of the graph. In this way the proposed adjustments infrastructure ski resorts can be integrated into the simulation. Improvements must, however, be added to the system so that the simulation is complete, especially at the influence of bars and restaurants that can change behavior locally in the graph.

4 Conclusion

The model Ski-Optim was tested for the ski area of Verbier and gave convincing results. It corresponds to the measures on the field. There is very little deviation between the counting at the turnstiles and the simulated values for each time slot. An emergence of the spatial structures of the ski runs congestion is not only shown by the model, but confirmed by the observation. It is possible to test the skiing infrastructure (for example the change of a ski lift to a chairlift or a new ski lift) and to see simply the effects on the skier flow. From the methodological point of view, the Ski-Optim model can be transposed efficiently on all issues on networks with queues or jams such as urban traffic. Furthermore, this approach has the advantage to give the opportunity to handle an unknown or only partially known situation. Some developments are in progress, such as the linkage of the variation of the cableway rate of flow depending on the forecast of the queues, in order to reduce the energetic impact of the cableways.

References

[1] JC Loubier, Les sports d'hiver en mutation, chapter 6: Le changement climatique comme facteur de mutation des pratiques sportives de masse, ;,in Bourdeau Ph, Les Sports d'hiver en mutation Coll finance. gestion. management, 89-97 2007.

[2] R. W. Wolff, Stochastic modeling and the theory of queues (1989). (Vol. 14). Englewood Cliffs, NJ: Prentice hall.

[3] M Revilloud, J-C Loubier, M Doctor, M Kanevski, V Timonin and M Schumacher, Predicting Snow Height in Ski Resorts using an Agent-based Simulation (2013), in: *Multiagent and Grid Systems* (MAGS). 9(4);279-299

[4] M Wooldridge; An Introduction to Multiagent Systems - Second Edition, John Wiley & Sons. , 2009:

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Real-time detection of anomalous paths through networks

Steven D. Prager
University of Wyoming
Department of Geography
Laramie, USA
sdprager@uwyo.edu

R. Paul Wiegand
University of Central Florida
Institute for Simulation & Training
Orlando, USA
wiegand@ist.ucf.edu

Abstract

The proliferation of increasingly inexpensive mobile devices capable of transmitting accurate positional information to other devices and servers has led to a variety of applications ranging from health situation monitoring to GPS-based offender monitoring. One of the resultant challenges is in understanding, in real-time, when incoming observations merit further examination. In this research, we investigate an approach for identifying anomalous paths through networks using real-time comparisons to a previously learned model. Our approach, the development of a series of "posterior weighted graphs" allows us to both determine which underlying model a particular path most closely represents as well as evaluate this relationship in real-time as more observations become available. Here we present the posterior weighted graph approach for examining path similarity and an extension for detecting anomalies in real-time. Our results illustrate how we can distinguish from among multiple candidate paths and, likewise, when observations no longer match an expected model.

Keywords: path similarity, anomaly detection, networks, mobility, GPS

1 Introduction

One challenge in understanding paths through networks is detecting when observed paths depart from what is considered normal. What is normal is, of course, subject to *a priori* establishment of corresponding expectations. In this paper we present an approach for learning an *a priori* model for a set of potential paths. We then demonstrate how this model can be used to facilitate real-time detection of when observed paths depart from the expected path(s) represented by the learned model. Applications of real-time path anomaly detection range from the health field [7] to fraud detection [3], to automated surveillance of individuals, traffic, objects and crowds [9].

In the context of this study, we define anomalous event as an event that has characteristics significantly different than normal [9]. Proliferation of track data from mobile devices has led to a variety of applications wherein the goal is to detect anomalous mobility patterns [2, 7]. In such cases, the anomaly occurs when an observed mobility pattern departs from a previously established pattern. Often couched in terms of "path matching" problems [6], many methods are used to look at path similarity [4].

The challenge of working with similarity detection methods for real-time path data is that the paths and, hence, the corresponding metrics are constantly changing [5]. Similarly, there are a potential for a number of ambiguous cases [8]. Alternatively, it is possible to classify a dynamic path against an established baseline [1, 3]. In both [3] and [1], a baseline is established with previously collected GPS traces. Though [3] uses a grid-based approach in conjunction with isolation-based methods and [1] uses a reduced "support point" representation, both compare emerging trajectories to a previously established baseline.

Here, we present a method capable of using either previously collected GPS data or baseline paths from a map interface such as Google Maps. In turn, we present a new method for discerning departures from this baseline using a series of weighted graph models. In the next section we address the problem and, following, illustrate the methods and analytic results.

2 Problem Definition

The principle emphasis of this research is to determine whether an observed path departs from an expected path and to make this determination in real-time.

Consider a street network represented by a series of nodes and edges. Paths through that network can be represented as a collection of ordered vertices where, by extension, traversal of a vertex implies traversal of the corresponding edge between a vertex and the previous vertex. Paths may be thought of in terms of being either observed (i.e., a series of recorded network locations), or as expected (i.e., determined in an *a priori* manner).

Observed paths may be thought of in terms of whole or partial paths. Whole paths are simply paths between an identified origin and destination. Partial paths may be either a static segment of a whole path or a path that lengthens dynamically over time with or without a predetermined destination. For this effort we focus on the latter, paths that evolve over time with no predetermined destination. While any network space may be used, we express observed paths via serial latitude and longitude locations and, in turn, associate these observations with the nearest network vertices in a planar embedded street network.

Expected paths are determined in an *a priori* manner and represent idealized versions of paths that will be observed. Expected paths are a set of edges, specified via either previously recorded locations, algorithmically via a shortest path between two points, or manually via an appropriate interface).

In order to determine whether an observed path is departing from an expected path two assumptions are necessary. First, for a variety of reasons, an observed path through a network may deviate from what is expected but may still reasonably be considered to be the same (e.g., a parallel road used to divert around an obstruction in a street network). Thus, the basis for determining when an observed path has substantively departed from what is expected must be couched in terms appropriate to the problem at hand.

Allowance for relative path similarity is accomplished through the establishment of a "decay" function around the expected path. This decay function serves to distribute the highly discrete information associated with a specific path on to adjacent edges in an exponentially declining manner relative to the cumulative shortest-path distance to each node encountered in the expected path. We call this representation a "posterior weighted graph" (PWG) and it is the model against which observations are compared.

Identification of departures of observed data from expected paths in real-time also requires the establishment of lower bound criteria for when an observed path is no longer functionally equivalent to an expected path. This lower bound is determined by two parameters, the maximal rate of change of observed data relative to the expected path models, and a threshold time in which no new maxima occur.

In the next section, we formalize the modelling approach. We briefly describe the development of the posterior weighted graph models, the classification process used to compare observations to expectations, and our real-time implementation of this process.

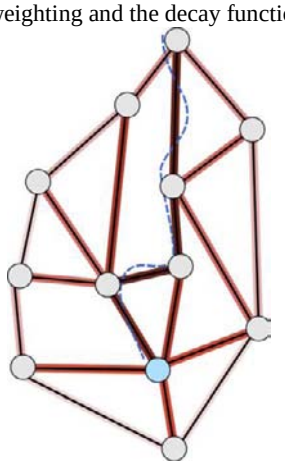
3 Detection of path anomalies

3.1 Characterization of expected paths

As mentioned in the previous section, expected paths are characterized using posterior weight graphs. The PWG probabilistically represents the likelihood that any edge will be used in association with an expected path.

The PWG is created by first initializing every edge in the graph with a 0 weight. The vertex sequence associated with the expected path is then traversed, and the coincident edges are each assigned an initial weight value. Following the assignment of the initial weight value (usually the edge length, but any weight may be used), the edges in the neighbourhood of each vertex are then assigned progressively lower weights using $e^{-dist(i,j)/\sigma^2}$ where *dist(i,j)* is the cumulative shortest-path distance to the next vertex or vertices, and σ is the decay parameter. The depth of the neighbourhood traversal is limited by parameter T_w, a threshold weight tolerance below which the decay is considered to render edge weights negligible in terms of their influence on the model.

Figure 1: Edge weighting and the decay function.



As Figure 1 illustrates, the edges associated with the expected path (dashed line) are most strongly weighted. The edges immediately adjacent are weighted somewhat less strongly, and distant edges are weighted in a very limited manner. The process of traversing vertices in the path is repeated until the expected path is complete. Because the neighbourhood of each vertex is examined, edges coincident to multiple vertices are reinforced.

3.2 Classifying a path

3.2.1 Converting PWGs to a probability model

As there may be multiple PWGs for multiple expected paths, it is necessary to set the stage for modelling any observed path as a set of edge probabilities. This supports a classifier that uses a probabilistic approach to determine from which expected path a set of observed edges would be most likely drawn. The probability model for each expected path is derived from the corresponding PWG.

We begin the process by establishing a minimum edge probability, p_{min}, an arbitrarily low probability that ensures that no edge has zero probability. We then rescale all the edge weights based on the maximum weight less p_{min} and add p_{min} to all of the probabilities (Eqs. 1 and 2).

$$w_{max}^{k} = \frac{max_{w \in W_k} w}{(1.0 - p_{min})} \quad (1)$$

$$p_{ij}^{k} = \frac{w_{ij}^{k}}{w_{max}^{k}} + p_{min} \quad (2)$$

This scaling process is repeated for each *k* expected paths and ensures all weighted edges from the PWGs have some minimum probability, that the weight values are monotonically proportional to edge probabilities, and that all potential expected probability models are scaled to the same p_{min}.

3.2.2 The anti-model

Determining when observations depart from expectations as represented by the set of expected probability model requires an additional mechanism. Specifically, as the classifier will identify the probabilistically "best" match even if the corresponding probabilities are very low, we must provision for the case when the path being classified does not strongly match any of the individual expected path probabilities. In order to facilitate this process we develop what we refer to as the "anti-model."

The anti-model is essentially a reciprocal set of probabilities associated with edges not reinforced by the *k* expected models. First, for each edge in each of the *k* expected models, the maximum probability for that edge is determined. The anti-model probability is, in turn, calculated for each edge as the minimum of either the complement of the maximum probability or a user-defined parameter, $p_{sensitivity}$ (Eq. 3).

$$p_{i,j}^{anti-model} = min\{1 - p_{i,j}^{max}, p_{sensitivity}\} \quad (3)$$

This results in a final probability model wherein edges with high probabilities in any of the *k* expected models are assigned a low probability through the $p_{sensitivity}$ parameter. This mechanism implements a heuristic, worst-case identification of an anomalous path. Such a worst-case identification reduces the possibility of falsely identifying anomalous paths. We now explore how this is used in the classification process.

3.2.3 Classifying an observed path

Given the probability models associated with each expected path and the corresponding anti-model, we wish to determine whether an observed path is most like one of the *k* expected paths or most like the anti-model.

In order to do to this, we compute the log likelihood of the observed path being from any given model (Eq. 4). This represents an assessment of the likelihood of the joint event that the edges in the path set came from model *k* under the assumption that edge inclusions are conditionally independent given the model.

$$Pr\{path \text{ from model } k\} := \sum_{(i,j)\in path} \log p_{ij}^{k} \quad (4)$$

It is unlikely that the assumption of conditional independence is completely valid. Nevertheless, we believe that the graph contains sufficient information so that proceeding with the naïve assumption still results in a useful classifier.

Finally, for classifying paths, we will typically include the anti-model in addition to the *k* expected models. After all log likelihoods have been calculated, the model with the highest log likelihood is the model from which the observed path is most likely drawn. If, on the other hand, the anti-model has the highest log likelihood, then we assert that the observed path does not likely match any of the expected paths and can, therefore, be considered anomalous.

3.3 Real-time detection of anomalies

Once the classifier is established, extending it to work with real-time observations is relatively straightforward. Simply, we consider a path to have a starting observation and, over time, successive additional increments of the path are added. In terms of classifying a dynamic set of observations, for each successive observation, the cumulative "observed" path is extended and the classifier is reapplied relative to the original expected models and corresponding anti-model. The challenge is detecting when a set of observations has transitioned from an expected state to an anomalous state.

In order to detect transitions from expected to anomalous in real-time, we use a second order numerical approximation of the backwards difference technique (Eq. 5).

$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{\Delta x} \quad (5)$$

For each additional observation (extension to the path), we instrument the real-time classifier to record the log likelihood for the each of the *k* models and the anti-model. When the trend with the highest log likelihood simultaneously expresses a maximum positive rate of change, we consider this a trigger (indicating the potential for association of the observations with a corresponding model). When the log likelihood for that trend does not decrease for a user specified number of additional "lock in" observations (L_o) and there are no additional triggers, the observed trajectory is considered to be similar to the corresponding model. If this is one of the expected models, then the observed data are considered expected, if the lock-in is associated with the anti-model then the observations are considered anomalous.

3.4 Summary

As with any modelling effort, the success of the model is dependent on the proper selection of the parameters underlying the model. The advantage of having an adequate parameter space, however, is that the model can be tailored to multiple modelling scenarios. For example, while our case studies use spatially embedded transportation networks (and have the commensurate topological constraints), the parameters would allow for use of other networks such as telecommunications networks, social networks, and utility networks. For the scenarios that follow, Table 1 summarizes

Table 1: Model parameters and description.

Parameter & Value	Description
σ =20.0	The rate of decay of edge weights associated with the model for each path. Larger values result in a more general model.
T_w = 0.00001	The weight tolerance controlling the depth of the decay function.
p_{min} =0.0001	Minimum probability for rescaling edge weights from decay model into probability model.
$p_{sensitivity}$ =0.2	A lower bound to limit false positive associations with the anti-model.
L_o = 3	Lock-in. This is the number of post-trigger observations required to confirm association with either an expected model or the anti-model.

the parameter space. In the present experiment, the parameter values were empirically identified and work across a variety of scenarios and input data. Future research will examine how appropriate parameter values can be derived through machine learning based on input training data.

In the next section we illustrate the use of the above model and demonstrate its use for both real-time path matching and anomaly detection.

4 Implementation and evaluation

4.1 Two scenarios

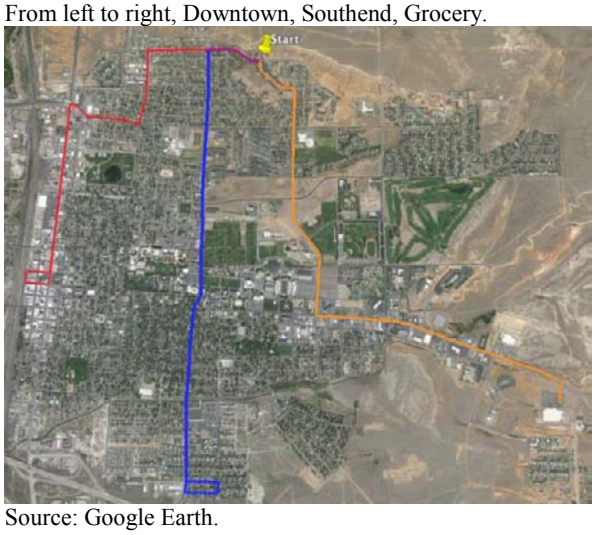
In order to demonstrate the ability of the previously described model to both identify when an observed path matches an expected path and when an observed path becomes anomalous, we present two related scenarios.

- Scenario 1: Multiple expected paths are learned and the anti-model is computed. Observed data are monitored and the point at which the observations are definitively associated with one of the expected paths is reported.

- Scenario 2: Multiple expected paths are learned and the anti-model is computed. Observed data are monitored and the point at which the observations can definitively be considered anomalous is reported.

The scenarios are based on a subset of the street network from Laramie, WY, USA with expected data based on routes derived from Google Maps and observed data collected using a Garmin Forerunner 210 GPS watch.

Figure 2: The learned paths shown on the Laramie streets. From left to right, Downtown, Southend, Grocery.



Source: Google Earth.

Both of the scenarios classify against three learned models and the anti-model. The three models include "Downtown," round trip travel to the Laramie town centre, "Southend," an arbitrary trip across town and, "Grocery," a trip to the grocery store. All of the paths were mapped in Google Maps, extracted as GPX data, and mapped to coincident vertices in the street network using a spatial search algorithm. The resultant ordered vertices are the graph-based representation of the potential expected paths.

Though the observed data were collected as a single GPS track, we simulate real-time online processing. The real-time emulation is accomplished by introducing each successive track point and extending the observed path. We then re-compute the log likelihoods, recalculate the numerical approximation of the second derivative, and evaluate against L_o.

4.2 Scenario 1 – Multiple expected paths

In this scenario we simulate where an individual is choosing from among several potential activities as specified in Section 4.1. Our goal is to observe their trajectory and, as quickly as possible, identify which activity they are most likely doing.

As Figure 3 illustrates (and can be intuited from Figure 2), it is not possible to differentiate the activity based on the initial set of observations. However, beginning with the fourth observation (49 seconds into the journey), the likelihood of any given path begins to diverge. The first trigger (maximum positive rate of change associated with the highest log likelihood) occurs at the 6th observation (81 seconds), and the association with the Southend route is locked in at the 9th observation, approximately 51 seconds later.

Figure 3: Confirmation of the Southend path at 132 seconds.

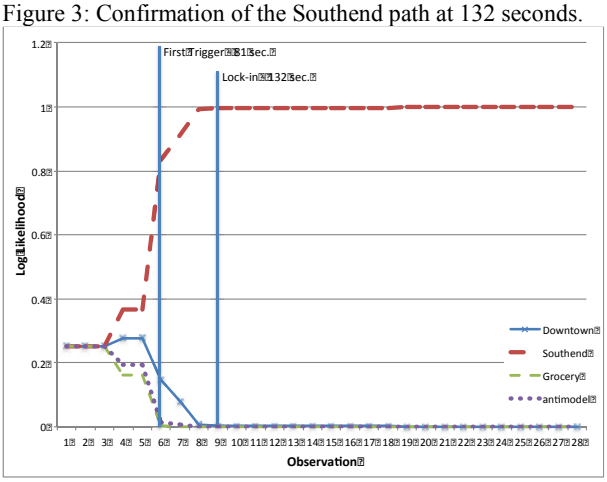
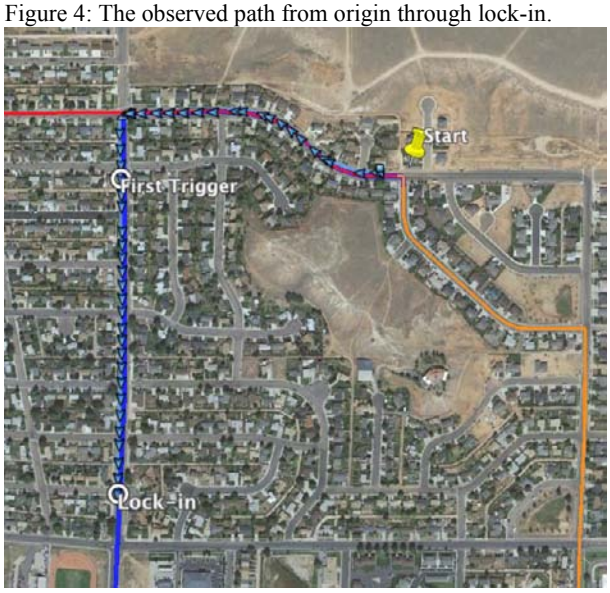


Figure 4 shows a map view of the first trigger and the subsequent lock-in. The "soft" association that comes from the trigger event helps minimize false positives and serves to leverage the fuzziness (and the potential that observations may match, depart, then return to a specific expected path) facilitated with the underlying decay function described in Section 3.1.

Figure 4: The observed path from origin through lock-in.

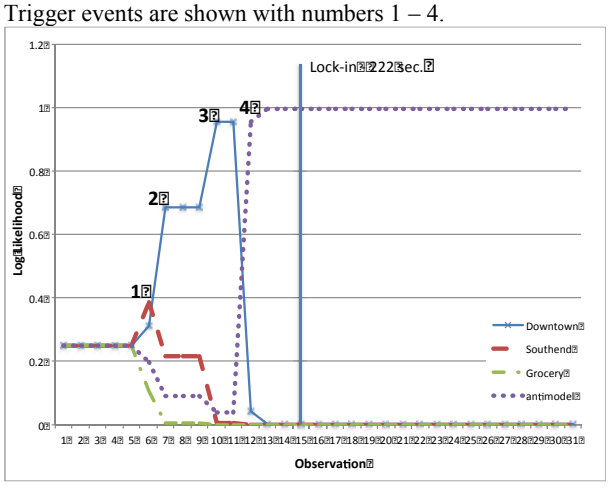


Source: Google Earth.

4.3 Scenario 2 – Expected vs. anomalous paths

Like the previous example, here we simulate a scenario wherein we are trying to determine if an individual's trajectory through the network is consistent with one of three predetermined paths. In contrast, however, rather then reporting when an individual's trajectory is associated with an expected path, we want to report when their trajectory is definitively anomalous.

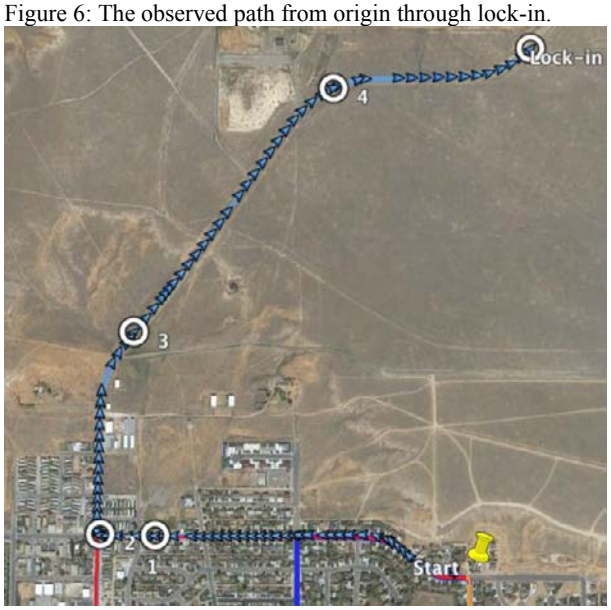
Figure 5: Confirmation of anomalous route at 222 seconds. Trigger events are shown with numbers 1 – 4.



In contrast to the previous example, the first trigger event in this case arises from an apparent association with the Southend path at 79 seconds (Figure 5, trigger 1). As illustrated, however, this is something of a false positive, and the lock-in fails with a second trigger event at 96 seconds in association with the Downtown path. This association is relatively strong, however, a third trigger event on the same model occurs with the 10th observation at 142 seconds. This third trigger event prevents the lock-in that would have otherwise occurred at this observation. At the 12th observation (174 seconds) a forth trigger event, this time associated with the anti-model, is seen. Three observations later (per T_w), there are no additional triggers and the lock-in as an anomalous path is confirmed at 222 seconds.

Again, Figure 6 shows a map view of the observed trajectory and the various detection events.

Figure 6: The observed path from origin through lock-in.



Source: Google Earth.

The sequence of triggers illustrates the role of the interacting decay functions in terms of defining the probabilities of associating with any given path. Since the probability of the observed data is cumulative in nature, there is a seeming lag between the path association and the trigger point. This is a characteristic of the approach and can be adjusted through the sensitivity and σ parameters.

4.4 Summary

In the presented scenarios, the paths themselves and the corresponding GPS data can clearly be differentiated from one another and the underlying anti-model using the presented method and corresponding parameters.

In a different context (e.g., that such as illustrated in [3]), the same approach could be used to determine whether a single GPS track is more like a collection of potential paths or, again, the anti-model. A characteristic of this approach is its flexibility supporting either comparisons to specific, individual paths or, alternatively, a collection of paths traversing the network in question. The learned anti-model can be the "reciprocal" of a single path or a collection of paths or segments. The application in question will be the key driver in decisions regarding the overall representation, definition of path start and end points, and whether or not specific, individual paths or path sections need to be identified.

5 Conclusion

This paper presents a preliminary method for using a classification-based approach for real-time interpretation of network observations. The presented approach is useful for discerning either when a set of observations is most similar to an expected path or unlike any *a priori* specified expectations. This latter case is useful for identifying anomalous paths in real-time.

The ability to detect either path similarity or difference is predicated on learning the model or models that characterize expected data. These models, along with the anti-model must be learned in the context of the specific problem at hand, the nature of the corresponding network, and the characteristics of the observed data. The parameters, while perhaps numerous, allow for the approach to be tailored to a variety of scenarios. While the presented approach is on a street network, any network with the potential for supporting expected and observed paths is a candidate for use with this method as the entire process is aspatial and based on network measurements and network locations.

Two key areas merit additional research. First, as previously mentioned, it would be useful to be able to learn the parameter space for different problem classes. This would enable more effective parameter selection depending on problem and network characteristics. The second area for additional research is in terms of improving the approach for handing real-time data. Predictive methods from the signal process and machine learning communities may prove very useful in this regard.

References

[1] J. A. Alvarez-Garcia, J. A. Ortega, L. Gonzalez-Abril, and F. Velasco. Trip destination prediction based on past GPS log using a hidden markov model. Expert Systems with Applications, 37(12):8166–8171, 2010.

[2] S. Buthpitiya, Y. Zhang, A. K. Dey, and M. Griss. N-gram geo-trace modeling. In Proceedings of the 9th International Conference on Pervasive Computing, Pervasive'11, pages 97–114, Berlin, Heidelberg, 2011. Springer-Verlag.

[3] C. Chen, D. Zhang, P. S. Castro, N. Li, L. Sun, and S. Li. Real-time detection of anomalous taxi trajectories from GPS traces. In Mobile and Ubiquitous Systems: Computing, Networking, and Services, pages 63–74. Springer, 2012.

[4] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, SIGMOD '05, pages 491–502, New York, NY, USA, 2005. ACM.

[5] Z. Fu, W. Hu, and T. Tan. Similarity based vehicle trajectory clustering and anomaly detection. In Image Processing, 2005. ICIP 2005. IEEE International Conference on, volume 2, pages II–602. IEEE, 2005.

[6] Q. Lu, F. Chen, and K. Hancock. On path anomaly detection in a large transportation network. Computers, Environment and Urban Systems, 33(6):448 – 462, 2009.

[7] T.-S. Ma. Real-time anomaly detection for traveling individuals. In Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility, Assets '09, pages 273–274, New York, NY, USA, 2009. ACM.

[8] F. Porikli. Trajectory distance metric using hidden markov model based representation. In IEEE European Conference on Computer Vision, PETS Workshop, volume 3, 2004.

[9] A. A. Sodemann, M. P. Ross, and B. J. Borghetti. A review of anomaly detection in automated surveillance. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 42(6):1257–1272, 2012.

Session:
Best Papers

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Linking crowdsourced observations with INSPIRE

Stefan Wiemann
Technische Universität Dresden
Geoinformation Systems
Dresden, Germany
stefan.wiemann@tu-dresden.de

Lars Bernard
Technische Universität Dresden
Geoinformation Systems
Dresden, Germany
lars.bernard@tu-dresden.de

Abstract

The combination of spatial data from the variety of sources on the web, being either legislative, commercially or voluntarily driven, is a major requirement for the establishment of a fully integrated geospatial web. Therefore, spatial data fusion techniques need to be linked to current web-developments, in particular on Spatial Data Infrastructures and the Semantic Web, to allow for standardized and effective use of combined spatial data for information retrieval. In this paper, crowdsourced environmental observations, representing the rapidly increasing amount of voluntarily collected data on the web, and INSPIRE, acting as the legal framework for spatial environmental information in Europe, are chosen to design and develop capabilities for spatial data fusion on the web. Possible use cases show mutual benefits for both volunteers and INSPIRE data providers, and might thus facilitate further collaboration. A prototypical implementation based on common SDI and Linked Data standards demonstrates the feasibility of the proposed approach and offers a starting point for further exploration.

Keywords: Data Fusion, SDI, Linked Data, Crowdsourcing, INSPIRE.

1 Introduction

The rapid development of location-enabled mobile devices and applications considerably influenced the public awareness, availability and use of spatial data in the recent years. Especially the amount of voluntarily collected spatial data on the web, often denoted as Volunteered Geographic Information (VGI, [4]), is continuously increasing. One of the most prominent examples is the OpenStreetMap[1] project for volunteered topographic mapping [5]. Furthermore, a large number of smaller projects, such as environmental monitoring campaigns [9], are establishing a versatile and most up-to-date basis for information retrieval and decision making that has not existed before.

In parallel, the INSPIRE (Infrastructure for Spatial Information in the European Community) regulation lays down the requirements for the unified and harmonized provision of spatial environmental data across the European Union und thus, serves as an ideal spatio-temporal reference for crowdsourced observations. In general, the integrated use of crowdsourced and administrative data will provide mutual benefits.

To link and combine arbitrary data sources on the web, service-based data fusion techniques need to be applied, independent from underlying data formats and provision means. Standards for information exchange and interlinking need to be established to support the interoperable and flexible orchestration of fusion processes and the effective combination of spatial data, one of the major building blocks towards a fully integrated geospatial web.

The paper introduces possible use cases for the fusion of crowdsourced environmental observations with INSPIRE reference data (chapter 2), followed by a short introduction to spatial data fusion (chapter 3). Subsequently, the requirements (chapter 4) and a prototypical implementation (chapter 5) for service-based spatial data fusion are described. Finally, a conclusion and outlook for further research is given (chapter 6).

2 Possible Use Cases

From the crowdsourcing perspective, INSPIRE can serve as a fundamental basis for validating environmental information on a broad range of topics, such as land cover and land use, environmental monitoring, habitat information or species distribution. However, INSPIRE will only be able to deliver coarse information with respect to spatial and (especially) temporal resolution, because of limited resources on the part of administrations responsible for data collection and provision. Thus, the use of INSPIRE data might be insufficient for data-intensive or real-time applications. The Eye on Earth[2] initiative already addresses this issue and tries to engage citizens to explore, collect and share environmental information in their surrounding and thereby complement existing data sources on the European level. However, developments to combine both crowdsourced and administrative data for information retrieval are still in the early stages. From the INSPIRE perspective, we identified

1 http://www.openstreetmap.org

2 http://www.eyeonearth.org

four different application use cases for the fusion with crowdsourced data:

1. *Data densification* can be applied to refine the response of a data service provider to a user request on a specific environmental phenomenon. At first, INSPIRE data is searched for information matching the request. If the data found is not yet sufficient in terms of spatio-temporal resolution or thematic attribution, crowdsourced data is selected and combined with the previous result to enhance the response accordingly. The data is provided to the user including lineage and quality information on the input and output data.
2. *Data enrichment* affects crowdsourced data that is not directly in the scope of the INSPIRE Annexes, but in any way connected to it. A user requesting environmental data automatically gets hints on related crowdsourced data, to indicate additional information sources. Moreover, relations to additional, explanatory or extending, data sources on the web facilitates the usability and applicability of INSPIRE data in general.
3. *Data update* describes the ability to refresh INSPIRE data with the help of related crowdsourced observations. Since data providers usually have limited resources for regularly updating information, crowdsourced information can give hints on where data actually needs to be updated. Therefore, both can systematically be compared to identify missing, incomplete, outdated or erroneous parts. The data provider is thus able to specifically investigate differences and, if applicable, update accordingly.
4. *Statistical analysis* of crowdsourced observations can benefit from the combination with INSPIRE, which acts as a persistent spatio-temporal reference system. Thereby, INSPIRE facilitates the comparability and ability to analyse crowdsourced information across the European Union.

All of the mentioned use cases require data fusion techniques to be applied, especially for matching, interlinking and resolving spatial data on the web. Furthermore, the awareness for lineage and quality information is considered crucial, because crowdsourced data will rarely comply with common INSPIRE quality standards.

3 State-of-the-art in data fusion

Within geosciences, the term fusion is quite ambiguous and frequently used within remote sensing, database research and spatial data processing [1]. With respect to signal processing, we hereinafter distinguish between data and sensor fusion. While sensor fusion is defined as the synthesis of multiple sensor measurements to receive comprehensive data on an observed phenomenon or entity [8], data fusion describes the combination of spatial data from multiple sources to provide a combined view that contains the most valuable data from the inputs [12]. Valuable hereby depends on the application context and purpose. Although sensor fusion is recognized as important for crowdsourced data, especially regarding consolidation and validation purposes, it is not in the scope of this paper. Instead, the focus is on data fusion, a task sometimes also referred to as conflation or data integration [10].

The classification of spatial data fusion is usually based on the operation direction (horizontal, vertical or temporal), input source (raster or vector), semantic level (representation, schema or ontology) or frequency (unique, periodic or real-time) [10, 11, 14]. The decision on a suitable fusion process depends on the application purpose, which can be change, discrepancy and error detection, data update and enrichment or the full integration of multiple spatial datasets [14]. Other factors describing a process can be supported in- and output formats, precision and recall rates or the computational performance.

To meet the requirements of web-based and standard compliant fusion of spatial data, a service-based approach is aspired. Therefore, complex fusion processes need to be decomposed into well-defined atomic processes in order to match the requirements of a Service Oriented Architecture (SOA) [3]. Here, we extend the previous approach presented in [13] and introduce the seven sub-processes shown in Figure 1 for the classification of service-based spatial data fusion. However, implementations might fall into more than one category and certain components may be optional, iterated or concatenated in different ways. Thus, the classification can be seen as an abstract reference framework for fusion processes.

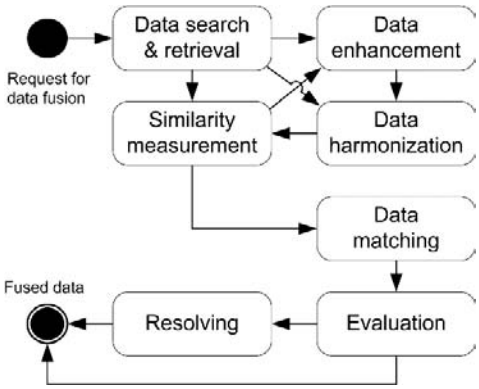
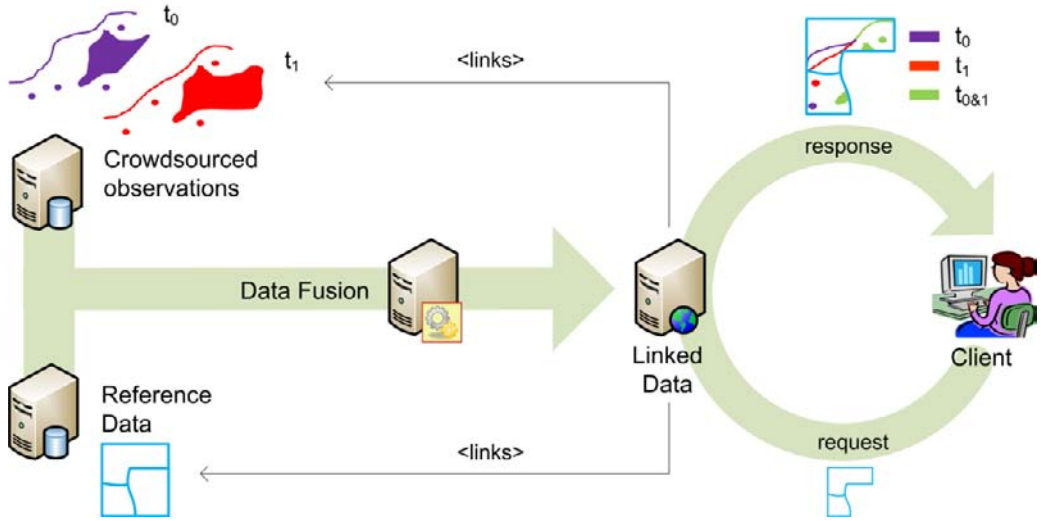


Figure 1: Classification for service-based spatial data fusion

4 Technical fusion aspects

For implementing web-based spatial data fusion, a number of service components for data provision, processing and search are required. For data provision and retrieval, INSPIRE specifies Data Download Services to be implemented for each data specification and member organization. The corresponding Technical Guidance currently obligates the use of the OGC (Open Geospatial Consortium) Web Feature

Figure 2: Generalized data fusion workflow for the combination of crowdsourced observations and spatial reference data



Service (WFS), but further specifications, such as the OGC Sensor Observation Service (SOS), the Web Coverage Service (WCS) or Linked Data are feasible [7]. For crowdsourced observations, the use of OGC services is desirable, but cannot be obliged. Communities often tend to develop their own data models and interface specifications to fit their specific purpose and thus, requires additional efforts on harmonization within the INSPIRE context. The applied services for spatial data processing should support the flexible and interoperable application and exchange of data fusion functionality across the web. Here, the use of the OGC Web Processing Service (WPS) is encouraged to comply with open standards. For the search of spatial data, the corresponding specification obligates the use of the OGC Catalogue Services for the Web (CSW) for INSPIRE [6]. For crowdsourced information, either the CSW or other open registries can be applied, as long as it can be accessed and requested online in a standardized manner.

The interlinking of crowdsourced observations with INSPIRE data can benefit from Semantic Web developments using Linked Data technology. Identified links can be encoded using RDF (Resource Description Framework) and either embedded directly in the crowdsourced data or managed as a standalone repository, which links to the corresponding data sources. Both approaches do not affect any INSPIRE infrastructure, but provide an additional layer for adding value to it. Beside the crowdsourced observations, further sources on the web, such as detailed descriptions on observed environmental phenomena by expert groups, can be linked as well. In addition, the service-based approach allows for the reuse of implemented data fusion functionality in other applications.

A generalized workflow for the fusion of crowdsourced and administrative data is depicted in Figure 2. Here, the starting point is a citizen collecting environmental information on a specific phenomenon in the field. This data is uploaded to a corresponding web portal, validated and stored accordingly. Depending on the application, this eventually triggers the fusion of the crowdsourced data with existing datasets, for which the selection is based on the observed phenomenon and its spatio-temporal extent. During the data fusion process, links between the input sources are generated and stored within a Linked Data triple store. Finally, a user requests and receives data on the phenomenon, including the option for added crowdsourced information. However, any of the use cases described in chapter 2 could be performed instead.

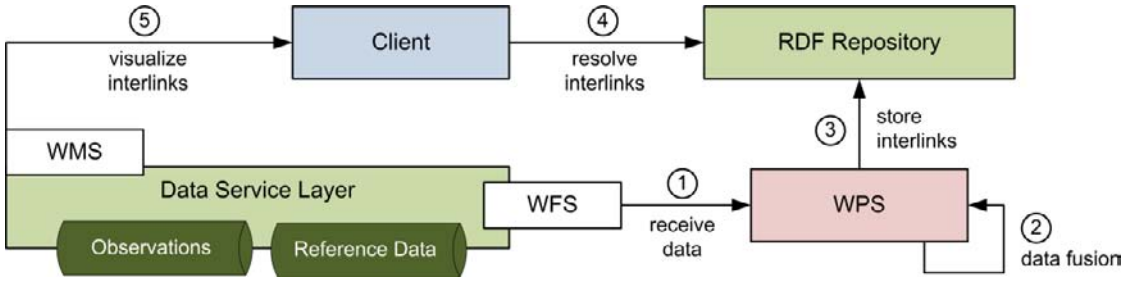
One of the most important aspects for data fusion as presented here, is the handling of quality information attached to the input sources, which determines the quality of a result and accordingly it's fit for purpose. Many applications or decision making processes require a certain level of data quality and thus rely on this information. To facilitate standardization and information sharing, quality associated to the input and output of a process should be formalized and encoded in a standardized format compliant to ISO 19115. While this is already mandatory for INSPIRE data, it is still a considerable challenge to add reliable quality information to crowdsourced data. The services for spatial data fusion must be capable of handling the quality information attached to the input sources. Hereby, the minimum requirement is to compile and keep track of the information from the input sources up to the final result. However, the optimum solution is a real utilization of quality information during the process, enabling propagated quality measurements based on input qualities and certain process characteristics.

5 Implementation prototype

The prototypical implementation for the fusion of crowdsourced environmental observations with administrative reference data is taken from the COBWEB project, which mainly deals with the development of on online-system for crowdsourcing in UNESCO Biosphere Reserves.

The current prototype manages the fusion of citizen observations on flora and fauna with INSPIRE relevant data on natural habitats. The basic workflow is depicted in Figure 3 and comprises the following steps:

Figure 3: Prototype workflow for the fusion of spatial data using OGC standards and a Linked Data repository



1. Crowdsourced observations and corresponding reference data is set up and provided via the OGC WFS interface for download. The fusion process, accessible via OGC WPS interface, requests this data as input for further processing.
2. The data fusion process is performed and relates the input features based on a number of similarity measurements, in particular bounding box overlap, Hausdorff distance and geometry buffer overlap for geometry objects and the Damerau-Levenshtein string distance for attribute comparisons.
3. All feature relations are encoded and stored as RDF triples including the corresponding WFS feature ids and underlying similarity measures.
4. The Client accesses the stored RDF triples and resolves the spatial features participating in a relation based on their feature id.
5. By using the OGC WMS interface on top of the data stores, the Client visualizes the resolved features by requesting a corresponding map overlay. A WMS GetFeatureInfo request can be generated to add detailed information on selected features.

The prototype currently follows a rather pragmatic approach to demonstrate the feasibility of the concept presented in this paper. To achieve the full potential of linking crowdsourced information to INSPIRE, further developments will focus on the support for all possible feature relations, the handling of quality information, automated orchestration of loosely-coupled fusion services with respect to the application use case and further interweaving the implementation with Semantic Web components.

6 Conclusion

So far, we demonstrated, that service-based spatial data fusion enables the combined use of multiple spatial data sources on the web for the retrieval of value-added information. Although still in the early stages, we can imagine a fully flexible, interoperable and versatile online system for dynamically interlinking and fusing any kind of spatial data on the web, with particular focus on the integration of the rapidly growing amount of voluntarily collected spatial data. Further developments and an evaluation of the approach will be documented in further publications.

The combination of INSPIRE data sources with data from crowdsourcing initiatives, offers great potential to create a comprehensive, most up-to-date and ubiquitously accessible source for environmental information in Europe. It combines the advantages of administrative data, namely quality assurance and the normative status, and crowdsourced data, with its rapid update cycle and partially high spatio-temporal resolution. The provided use cases show, that mutual benefits can be achieved from an advanced collaboration between both.

Still, one of the biggest challenges within the field of spatial data fusion remains the application-driven generation of value-added information from interlinked data. Therefore, it needs to be further analyzed how interlinked data can be selected and combined in an optimal fashion to serve a specific application purpose, and how quality information can be formalized and used to assist the fusion process. Although a lot of research questions still need to be solved, this will pave the way towards a Semantic Geospatial Web as proposed by Egenhofer [2].

References

[1] J. Bleiholder, F. Naumann. Data Fusion. *ACM Computing Surveys* 41(1):1-41, 2008.

[2] M. J. Egenhofer. Toward the semantic geospatial web. *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, McLean, USA, pp. 1-4, 2002.

[3] T. Erl. *SOA - Principles of Service Design*. Prentice Hall, 2008.

[4] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211-221, 2007.

[5] M. Haklay, P. Weber. OpenStreetMap: User-Generated Street Maps. *Pervasive Computing, IEEE* 7(4):12-18, 2008.

[6] INSPIRE. Technical Guidance for the implementation of INSPIRE Discovery Services. *Initial Operating Capability Task Force for Network Services*, Version 3.1, 2011.

[7] INSPIRE. Technical Guidance for the implementation of INSPIRE Download Services. *Initial Operating Capability Task Force for Network Services*, Version 3.1, 2013.

[8] OGC. OGC Fusion Standards Study Engineering Report. *OGC Public Engineering Report*. Open Geospatial Consortium, 2010

[9] H. E. Roy, M. J. O. Pocock, C. D. Preston, D. B. Roy, J. Savage, J. C. Tweddle, L. D. Robinson. Understanding Citizen Science & Environmental Monitoring. *Final Report on behalf of UK-EOF*. NERC Centre for Ecology & Hydrology and Natural History Museum, 2012.

[10] J. J. Ruiz, F. J. Ariza, M. A. Ureña, E. B. Blázquez. Digital map conflation: a review of the process and a proposal for classification. *International Journal of Geographical Information Science* 25(9):1439-1466, 2011

[11] A. Schwinn, J. Schelp. Design patterns for data integration. *Journal of Enterprise Information Management* 18(4):471-482, 2005.

[12] S. Stankuté, H. Asche. An Integrative Approach to Geospatial Data Fusion. *Computational Science and Its Applications – Proceedings of ICCSA 2009*, Suwon, Korea, pp. 490-504, 2009.

[13] S. Wiemann, L. Bernard. Conflation Services within Spatial Data Infrastructures. 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal. 2010.

[14] S. Yuan, C. Tao. Development of Conflation Components. *Proceedings of Geoinformatics'99 Conference*, Ann Arbor, USA, pp. 1-13, 1999.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Capability of movement features extracted from GPS trajectories for the classification of fine-grained behaviors

Ali Soleymani
Department of Geography
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich, Switzerland
ali.soleymani@geo.uzh.ch

E. Emiel van Loon
Computational Geo-Ecology, IBED
University of Amsterdam
PO Box 94248, 1090 GE
Amsterdam, The Netherlands
e.e.vanloon@uva.nl

Robert Weibel
Department of Geography
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich, Switzerland
robert.weibel@geo.uzh.ch

Abstract

Recent advances in tracking technologies provide an unprecedented opportunity for a better understanding of animal movement. Data from multiple sensors can be used to capture crucial factors deriving the behaviors of the animal. Typically, accelerometer data is used to describe and classify fine-grained behaviors, while GPS data are rather used to identify more large-scale mobility patterns. In this study, however, the main research question was to what extent fine-grained foraging behaviors of wading birds can be classified from GPS tracking data alone. The species used in this study was the Eurasian Oystercatcher, *Haematopus ostralegus*. First, a supervised classification approach is employed based on parameters extracted from accelerometer data to identify and label different behavioral categories. Then, we seek to establish how movement parameters, computed from GPS trajectories, can identify the previously labeled behaviors. A decision tree was developed to see which movement features specifically contribute to predicting foraging. The methods used in this study suggest that it is possible to extract, with high accuracy, fine-grained behaviors based on high-resolution GPS data, providing an opportunity to build a prediction model in cases where no additional sensor or observational data on behavior is available. The key to success, however, is a careful selection of the movement features used in the classification process, including cross-scale analysis.

Keywords: Movement analysis, GPS, accelerometer, foraging behavior, movement parameters, classification

1 Introduction

Classification of movement trajectories into different behavioral categories has become a recent trend in many domains, including e.g. movement ecology, transportation, and urban management. In ecology especially, behavioral classification is an important analysis step, because knowledge about behaviour provides important input to many inferences about physiology, energy balance, and evolution of particular species. While various types of data are being used for animal behavior classification, the use of features based on movement trajectories (e.g. GPS) is still quite uncommon (see [18]). The main reason for this has been that when the goal is to distinguish between behaviors (especially fine-grained behaviors, e.g. foraging vs. non-foraging), the temporal sampling rate is typically low or irregular in relation to the variability inherent to the movements that are considered. However, due to recent advances in tracking technologies, it has become feasible to collect high-resolution GPS and sensor data on a more regular basis. For example, GPS has been integrated into operational systems with other sensor technologies to collect temperature, activity, proximity and mortality data from terrestrial species and birds [1, 19, 21].

This study aims at developing a classifier to identify foraging behavior in a shorebird, the Eurasian Oystercatcher (*Haematopus ostralegus*), based on GPS trajectory data. This species has been intensively studied ([6]) to answer questions on e.g. foraging ecology, resource use and territoriality in shorebirds. The GPS trajectory data for individuals may be more accurate and less biased than the sighting or experimental data that are available from previous research and may thereby lead to more robust answers. Especially the time spent on foraging as well as foraging locations form important variables to measure foraging strategies and efficiency.

Accelerometer data can be used to identify various behaviors of an oystercatcher, including foraging [18], the same way as depth loggers are used to record 'dives', salinity sensors to record 'being in the water', or light sensors to record 'being in a burrow' [7, 9, 10, 13, 17]. However, accelerometers are not yet in widespread use today and a lot of trajectories with location-only information have been collected and will continue to be collected. According to Movebank (www.movebank.org) as one of the major repositories of animal movement, more than 90% of the data collected there is location-only. Therefore we attempt to develop features and a classifier that is based exclusively on location data. In order to do so, the model of [18] is first used to generate the behavioral labels and then serves as a baseline to train and evaluate the classification model that is based exclusively on movement features extracted from GPS trajectories. Thus, the main research question is to what extent fine-grained foraging behaviors, on the example of oystercatchers, can be classified from GPS tracking data alone.

2 State of the art

A variety of methods for inferring behaviors based on sensor data have been proposed. Among movement parameters computed from trajectories, velocity has been used to distinguish between traveling and resting during bird

migration [9], identification of different behavioral categories in combination with accelerometer readings [18], and distinguishing behavioral drug treatments in neuro-pharmacology [3]. A combination of velocity and direction has also been used in [20] for defining behaviorally consistent movement units. Sinuosity, on the other hand, has been used for detection of behavioral change in animal movement [16], foraging movement and activity patterns of seabirds [25], and for distinguishing between trajectories of different vehicles types [4]. Wavelet analysis has also been applied based on the values of net displacement [23] and velocity [15] for studying behavioral patterns in animal movement.

Accelerometer data, on the other hand, is increasingly being applied to characterize behavior or describe certain movements, e.g. of humans (using accelerometers on smart phones) [24], domestic animals [12], as well as free-ranging animals like birds [10, 14, 17, 18] and marine mammals [7, 13].

3 Methods

In this paper, we use a data set of combined GPS and accelerometer observations, obtained in the Dutch Wadden Sea, south of the island Schiermonnikoog on 12 individual Eurasian Oystercatchers (*Haematopus ostralegus*). The birds were tagged with UvA-BiTS devices [1], and samples from June and July 2009 as well as from May and June 2011 were used in this study. There were different sampling intervals in the samples, but for the major part of the data it was one location per 13 seconds (the second large group was with intervals of 6 seconds and the intervals were always lower than one location per 45 seconds).

We first classified the Oystercatcher trajectories as 'foraging' versus 'non-foraging' based on accelerometer data, using a classification model introduced in [18]. In [18], the model had been calibrated for the same species at approximately the same location while using the same devices. Based on the labeled data set we then started to develop features and classifiers based on GPS data only. The following (movement) features were calculated for each fix of the trajectories: distance traveled; velocity; turning angle and its dependent variables including angular velocity (turning angle over time) and meandering (turning angle over distance traveled). See [3] and [4] for some example uses of these parameters. Furthermore, two parameters indicative of path curvature were generated: sinuosity and the Multi-Scale Straightness Index (MSSI; see [16]).

A decision tree was selected for the classification process, using the implementation in RapidMiner 5, (RapidMiner, http://rapidminer.com/). A top-down procedure is applied based on the CART learner to traverse the tree [2]. Whenever a new node is created at a certain stage, an attribute is picked to maximize the discriminative power of that node with respect to the examples assigned to the particular subtree. This discriminative power is measured by the information gain ratio [2]. The information gain ratio can be considered as the importance of the selected attributes in the design of the tree. This was the reason for choosing decision trees in this study: they can give an insight into the relative importance of different movement features in the identification of behaviors, by their appearance as a node splitter. Other machine learning methods such as SVM might even result in a slightly better classification performance (as preliminary test have shown), but since improving the classification performance was not the main objective of this study, those classification methods were not chosen. A 10-fold cross-validation procedure was applied to see how good the resulting classification performances are when different movement parameters were used as input variables. For the evaluation of the performance of classification models, we looked at different criteria, such as overall classification accuracy and Kappa values, as well as precision and recall values in the case of individual classes, specifically when we examined the foraging class.

Since the sampling intervals differed between data sets and earlier studies had demonstrated the importance of scale in the computation of movement parameters, we performed a cross-scale analysis, employing the method proposed by [11]. Values of movement parameters for each fix of the trajectory were computed across a series of sliding windows with different sizes of w, in a segment where $w/2$ fixes exist before and after the central sample point of interest.

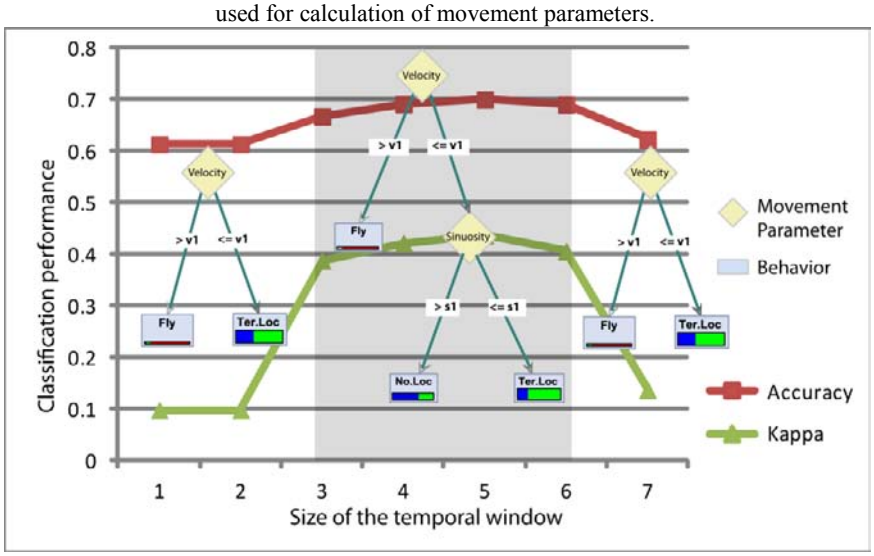
4 Results

4.1 Attribute selection

The classification performance was first acquired individually for all parameters. At first glance, velocity and distance traveled did seem to have a large impact on the classification results, which is in accordance with the findings of the studies having used these parameters [9, 15, 20, 23]. Turning angle, angular velocity and meandering, on the other hand, were not so helpful, which might be due to the positional error in GPS observations, especially at lower speeds. For the path curvature parameters, including MSSI and sinuosity, the values were computed across different scales. MSSI is inherently a multi-scale measure and similarly to sinuosity, it gives a ratio of the beeline distance between two points of interest and the actual distance traveled. However, the difference between the measures is that distance is computed multiple times, over a variety of scales for both temporal granularity and observational window [16]. We chose a granularity value of 2 and window sizes of 4, 8, 12, 16, 20 and 24, respectively. When individual sets of MSSI values were used, they were not helpful in distinguishing between classes, but as will be shown later, when geographic location is integrated (latitude and longitude), they do show a great potential in improving the results.

The same cross-scale approach was employed for sinuosity. The window sizes chosen for calculation of sinuosity start from the surrounding fixes (window size of 1), increasing up to 7 points before and after (1, 2, 3, 4, 5, 6, 7). Then, each set of sinuosity values computed at different scales were considered separately as input features in the classification, to see how the performance and the resulting decision tree would vary. We used the 3-class category (no locomotion, terrestrial locomotion and fly) of [18] in this part, as we wanted to investigate the importance of scale effects on a known model. In the subsequent process, however, the classification is only between foraging and non-foraging classes, by considering the outputs of the cross-scale analysis.

Figure 1: Variation of classification performance (Accuracy and Kappa) according to different temporal window sizes used for calculation of movement parameters.



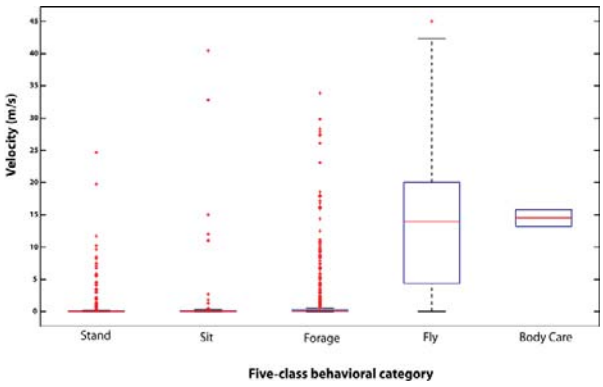
Interestingly, only after using a window size of 3 the role of sinuosity is starting to emerge in the structure of the decision tree (Figure 1). At the same time, for the window sizes of 3 to 6, higher classification accuracy and Kappa values were achieved. The tree structure for these window sizes was always the same, with velocity at the top, followed by sinuosity on the second level of the tree hierarchy (Figure 1). Since the window size of 5 scored relatively higher classification performance, it was selected as the window size at which sinuosity values can be reliably computed and considered as input features for the final classification.

4.2 Foraging versus non-foraging

In [18], a 5-class model has been calibrated that we are applying in this study; however we aggregate the output from 5 to 2 classes. First, since the fly class in the 5-class model can be easily distinguished from the stand, sit and foraging classes by using only the velocity parameter (Figure 2), the fly class is eliminated from the further analysis. The velocity values for the body care class are surprisingly high, which might be due to an error in the behavioral classification resulting from the accelerometer data. Nevertheless, since there were only two points labeled as body care, removing the fly class is still reasonable. Afterwards, all the non-foraging classes were aggregated and compared to the foraging class, resulting in a binary classification between a foraging class and a non-foraging class. Eliminating the fly class will help since there is a huge difference in the movement parameter values of the fly class and the rest of the classes, respectively, and if they were aggregated into a single class of non-foraging behaviors, it would have been difficult for the classifier to discriminate them. So, by first removing the fly class, only the sit, stand and body care classes will be aggregated into the non-foraging class. These behaviors share more similar movement characteristics.

In the end, there were 6486 fixes labeled as foraging and 4725 as non-foraging. Prior to applying the final classification, values of the selected attributes including distance traveled, velocity, sinuosity and MSSI are discretized into 3 bins, as it will help in improvement of the classification performance of the decision trees [5].

Figure 2: Boxplots of variation of velocity for five behavioral classes (Stand, Sit, Forage, Fly and Body Care).



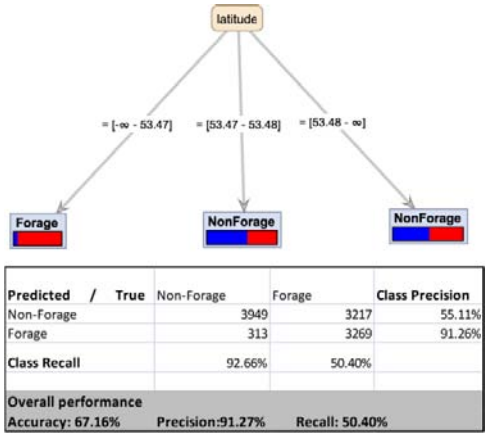
4.3 Importance of geographic context

To see whether knowledge about geographic context, represented by the geographic location of the birds, will help in identifying the behaviors, values of latitude and longitude of each fix were considered as input features in a classification tree. The resulting decision tree using only geographic location is shown in Figure 3. Apparently, latitude is a dominant variable in identifying behaviors, resulting in a

rather high classification accuracy of 67.16 %. However, very high classification precision (91.27 %) and at the same time very low recall values (50.40 %) does not indicate a robust performance. Nevertheless, this model will be considered as a baseline in order to compare with the following classification experiments, where values of movement parameters are integrated as well.

Subsequently, two separate decision trees based on the values of MSSI and sinuosity were developed (Figure 4). For each of these models, geographic location values were also integrated in order to make it possible to compare these to the baseline model developed in Figure 3. Additionally, since the importance of velocity and the distance traveled have been already emphasized, their values were also considered as input features in the classification model. Interestingly, both of the trees start with latitude at the top and then movement parameters are emerging at the lower levels (Figure 4).

Figure 3: The baseline decision tree for distinguishing foraging versus non-foraging developed based on location information, i.e. latitude and longitude. The confusion matrix is based on 10-fold cross-validation results.



Predicted / True	Non-Forage	Forage	Class Precision
Non-Forage	3949	3217	55.11%
Forage	313	3269	91.26%
Class Recall	92.66%	50.40%	
Overall performance			
Accuracy: 67.16%	Precision:91.27%	Recall: 50.40%	

5 Discussion

In the baseline classification model (Figure 3), the choice of latitude as a predictor variable in the decision tree can be understood from the east-west orientation of the Wadden island Schiermonnikoog, which provides the habitat of the studied individuals, located along the southern shore. The areas south of latitude 53.47º consist of mudflats with a short emersion time and high shellfish density. The area between 53.47º and 53.48º contains a combination of mudflats with long emersion time (which relates to a low shellfish density) and salt marshes. The area north of 53.48º contains salt marshes and meadow land. On the mudflats the Oystercatchers will feed on shellfish (mainly Baltic tellin – *Macoma baltica*) and ragworm (*Nereis diversicolor*). Conversely, on the saltmarsh and meadows they eat earthworms and insect larvae. The differences in habitat structure and prey types are reflected in different movement patterns.

As shown in Figure 4, the decision trees based on sinuosity and MSSI are not only improving the classification performance, but also give a more comprehensible overview of the importance of the movement features involved in combination with the underlying geographic location.

In the case of sinuosity, the leaves of the decision tree seem to be reasonable. Low values, indicating a smoother path, are labeled as foraging, whereas large values, indicative of a more complex path, are related to the non-foraging class (the path is more curved while the bird is sitting, standing or body caring due to GPS uncertainty). The values in the medium category are broken down again and distance values appear at the next level of the tree. The leaves at these levels are also sensible, as low and medium values of distance traveled are labeled as non-foraging and higher values as foraging. At the same time and as shown in Figure 1, it is worth noting that the usefulness of sinuosity is only revealed when the values are computed across different scales. In other words, if we had only used the sinuosity values computed at the original temporal rate, we could not have obtained the same results.

The resulting tree structure for MSSI is rather difficult to explain, but what looks interesting is the hierarchy in the structure of the tree (starting with window size 24x at the top and then 8x and 4x). Also, the tree is mostly dominated by foraging at the top (24x and 8x), while non-foraging only appears to be more dominant at the smallest scale (4x).

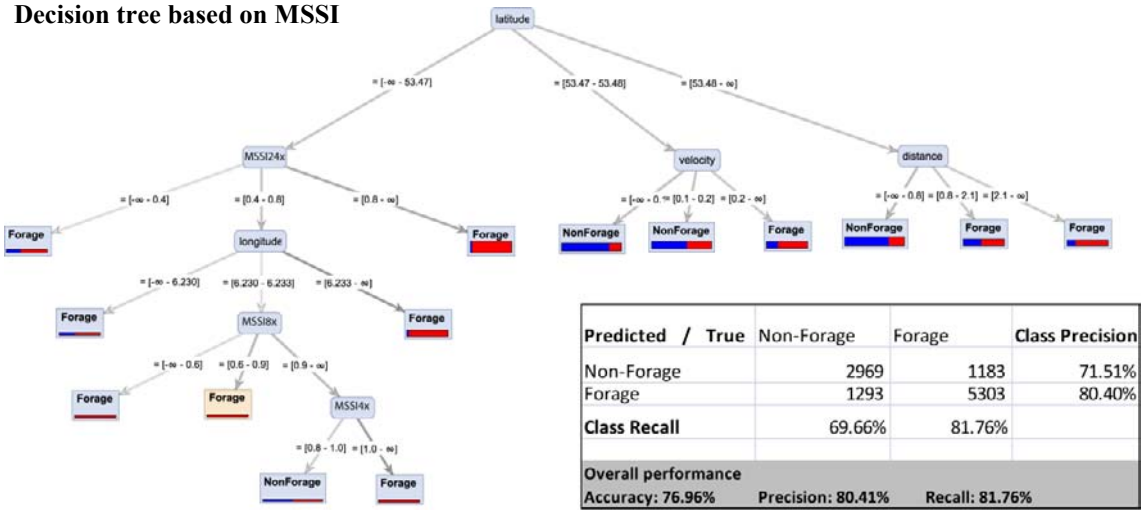
Resulting classification performances for the MSSI and sinuosity trees are comparable, with slightly better results for the sinuosity tree. As shown in the tables of Figure 4, overall accuracy and recall values are better for the sinuosity tree, whereas the MSSI tree results in a better precision value. Comparing to the baseline model developed based on geographic coordinates only (Figure 3), the classification performance is considerably better for the MSSI and sinuosity classification trees, leading to classifiers with an overall cross-validation accuracy of 0.78. This indicates a clear potential of parameters extracted from trajectories for the identification of movement-related animal behaviors.

6 Conclusions

To our knowledge, most of the works based on movement features (e.g. sinuosity and MSSI) do not use a classification model and are rather descriptive. Sinuosity, for example, has only been applied to flying birds ([8, 22]) and not yet to wading birds that are foraging on the ground. Thus, a classification model based on trajectory features, as presented in this study, seems a useful contribution to exploit information from animal-borne sensors to further understand and model animal behavior. However, apart from sinuosity and MSSI, there are other features that have not been used yet, including e.g. first passage time, scale invariance and fractal dimension. Exploration of these features can be considered as part of future work. Furthermore, since using GPS trajectory data often stumbles on problems with accuracy, an assessment of the positional accuracy and its consequences for the distinction of behavioral types seems important in order to fully appraise the potential of the proposed approach.

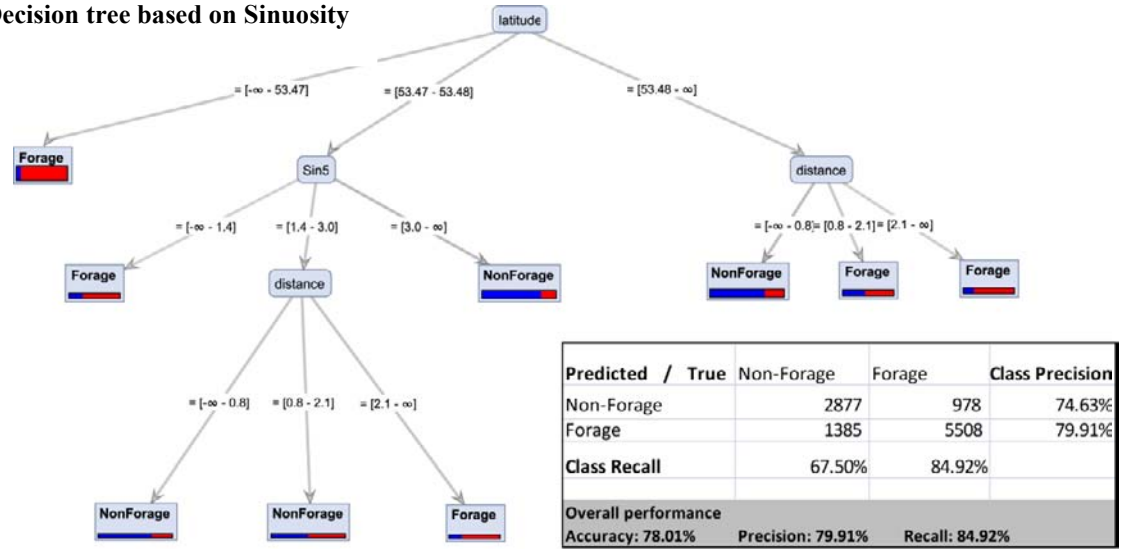
Figure 4: Two developed decision trees based on the two employed movement features (together with velocity and distance traveled): Sinuosity calculated at window size of 5 (shown as sin5) and MSSI calculated at window sizes of 4, 8, 12, 16, 20 and 24. Depending on their importance, each of these features are emerging at different levels of the corresponding decision trees. Note that the confusion matrices related to each tree are based on 10-fold cross-validation results.

Decision tree based on MSSI



Predicted / True	Non-Forage	Forage	Class Precision
Non-Forage	2969	1183	71.51%
Forage	1293	5303	80.40%
Class Recall	69.66%	81.76%	
Overall performance			
Accuracy: 76.96%	**Precision: 80.41%**	**Recall: 81.76%**	

Decision tree based on Sinuosity



Predicted / True	Non-Forage	Forage	Class Precision
Non-Forage	2877	978	74.63%
Forage	1385	5508	79.91%
Class Recall	67.50%	84.92%	
Overall performance			
Accuracy: 78.01%	**Precision: 79.91%**	**Recall: 84.92%**	

Acknowledgments

We gratefully acknowledge the participants of Dagstuhl Seminar 12512 for their contributions to the initial ideas for this project; Adriaan Dokter and the UvA-BiTS project for supplying the bird-tracking data (http://www.uva-bits.nl/); and COST Action IC0903 MOVE (http://www.move-cost.info) for funding part of this work.

References

[1] W. Bouten, E. W. Baaij, J. Shamoun-Baranes, and K. C. J. Camphuysen, "A flexible GPS tracking system for studying bird behaviour at multiple scales" *J Ornithol*, 571–580, 2013.

[2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and regression trees. *CRC press*, 1984

[3] J. Cachat, A. Stewart, E. Utterback, P. Hart, S. Gaikwad, K. Wong, E. Kyzar, N. Wu, and A. V Kalueff, "Three-dimensional neurophenotyping of adult zebrafish behavior" *PloS one*, 6(3): e17597, 2011.

[4] S. Dodge, R. Weibel, and E. Forootan, "Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects" *Computers, Environment and Urban Systems*, 33(6): 419–434, 2009

[5] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features" *In Proceeding of 12th International Conference on Machine Learning, Lake Tahoe, CA, Morgan Kaufmann, Los Altos, CA*, 1995, pp. 194–202.

[6] B. J. Ens, M. Kersten, A. Brenninkmeijer, and J. B. Hulscher, "Territory quality, parental effort and reproductive success of oystercatchers (Haematopus ostralegus)" *Journal of Animal Ecology*, 61(3): 703–715, 1992.

[7] A. C. Gleiss, S. J. Jorgensen, N. Liebsch, J. E. Sala, B. Norman, G. C. Hays, F. Quintana, E. Grundy, C. Campagna, A. W. Trites, B. a Block, and R. P. Wilson, "Convergent evolution in locomotory patterns of flying and swimming animals" *Nature communications*, 2: 352–358, 2011.

[8] D. Grémillet, G. Dell'Omo, P. G. Ryan, G. Peters, Y. Ropert-Coudert, and S. J. Weeks, "Offshore diplomacy, or how seabirds mitigate intra-specific competition: a case study based on GPS tracking of Cape gannets from neighbouring colonies" *Marine Ecology Progress Series*, 268: 265–279, 2004.

[9] T. Guilford, J. Meade, J. Willis, R. A. Phillips, D. Boyle, S. Roberts, M. Collett, R. Freeman, and C. M. Perrins, "Migration and stopover in a small pelagic seabird , the Manx shearwater Puffinus puffinus: insights from machine learning" *Proceedings of Royal Society*, 276: 1215–1223, 2009.

[10] A. Gómez Laich, R. P. Wilson, F. Quintana, and E. L. c. Shepard, "Identification of imperial cormorant Phalacrocorax atriceps behaviour using accelerometers" *Endangered Species Research*, 10: 29–37, 2009.

[11] P. Laube and R. Purves, "How fast is a cow? Cross-Scale Analysis of Movement Data" *Transactions in GIS*, 15(3): 401–418, 2011.

[12] P. Martiskainen, M. Järvinen, J. Skön, J. Tiirikainen, M. Kolehmainen, and J. Mononen, "Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines" *Applied Animal Behaviour Science*, 119: 32–38, 2009.

[13] Y. Mitani, R. D. Andrews, K. Sato, A. Kato, Y. Naito, and D. P. Costa, "Three-dimensional resting behaviour of northern elephant seals: drifting like a falling leaf" *Biology letters*, 6(2): 163–166, 2010.

[14] R. Nathan, O. Spiegel, S. Fortmann-Roe, R. Harel, M. Wikelski, and W. M. Getz, "Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures" *The Journal of experimental biology*, 215: 986–996, 2012.

[15] L. Polansky, G. Wittemyer, P. C. Cross, C. J. Tambling, and W. M. Getz, "From moonlight to movement and synchronized randomness: Fourier and wavelet analyses of animal location time series data" *Ecology*, 91(5): 1506–1518, 2010.

[16] C. M. Postlethwaite, P. Brown, and T. E. Dennis, "A new multi-scale measure for analysing animal movement data" *Journal of theoretical biology*, 317: 175–185, 2012.

[17] Y. Ropert-Coudert, F. Daunt, A. Kato, P. G. Ryan, S. Lewis, K. Kobayashi, Y. Mori, D. Grémillet, and S. Wanless, "Underwater wingbeats extend depth and duration of plunge dives in northern gannets Morus bassanus" *Journal of Avian Biology*, 40(4): 380–387, 2009.

[18] J. Shamoun-Baranes, R. Bom, E. E. van Loon, B. J. Ens, K. Oosterbeek, and W. Bouten. "From Sensor Data to Animal Behaviour: An Oystercatcher Example" *PloS one*, 7(5): e37997, 2012.

[19] J. Shamoun-Baranes, E. E. van Loon, R. S. Purves, B. Speckmann, D. Weiskopf, and C. J. Camphuysen, "Analysis and visualization of animal movement" *Biology letters*, 8(1): 6–9, 2012.

[20] A. Thiebault and Y. Tremblay, "Splitting animal trajectories into fine-scale behaviorally consistent movement units: breaking points relate to external stimuli in a foraging seabird" *Behavioral Ecology and Sociobiology*, 67(6): 1013–1026, 2013.

[21] S. M. Tomkiewicz, M. R. Fuller, J. G. Kie and K. K. Bates. "Global positioning system and associated technologies in animal behaviour and ecological research" *The Royal Society*, 365: 2163–2176, 2010.

[22] H. Weimerskirch, F. Bonadonna, F. Bailleul, G. Mabille, G. Dell'Omo, and H.-P. Lipp, "GPS tracking of foraging albatrosses" *Science*, 295: 1259–1259, 2002.

[23] G. Wittemyer, L. Polansky, I. Douglas-hamilton, and W. M. Getz, "Disentangling the effects of forage , social rank , and risk on movement autocorrelation of elephants using" *Proceedings of the National Academy of Sciences*, 105(49): 19108–19113, 2008.

[24] Z. Yan, Semantic Trajectories : Computing and Understanding Mobility Data, *PhD Thesis*. 2011.

[25] C. B. Zavalaga, G. Dell'Omo, P. Becciu, and K. Yoda, "Patterns of GPS tracks suggest nocturnal foraging by incubating Peruvian pelicans (Pelecanus thagus)" *PloS one*, 6(5): e19966, 2011.

POSTERS

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Visualization of uncertain catchment boundaries and its influence on decision making

Ulla Pyysalo
Finnish Geodetic Institute
P.O. Box 15, 02431
Masala, Finland
Ulla.Pyysalo@fgi.fi

Juha Oksanen
Finnish Geodetic Institute
P.O. Box 15, 02431
Masala, Finland
Juha.Oksanen@fgi.fi

Abstract

In this poster, we introduce an on-going project where uncertainty-aware drainage divides were calculated, visualized, and tested as background data for the decision-making process.

Keywords: Drainage divide, uncertainty, DEM, decision making.

1 Introduction

Societal decision making is often based on background information and its analysis, with a large portion of the information being spatial data. Common examples of this can be found in the fields of civil engineering, land use and transport planning, health care, and education. One objective of the Finnish National Geographic Information Strategy 2010–2015 is to use spatial information broadly in government decision making and to improve the political processes [1].

However, decisions are difficult to make based on uncertain data and models. Metadata reports on individual datasets are insufficient and do not effectively communicate the degree of uncertainty to users [2]. The uncertainty may take a variety of forms, such as errors, missing values, and deviations, which may originate from, for example, primary measurements, processing techniques, modeling, or interpolation. To ensure that the background data do not have a misleading impact on decisions, the characteristics of the underlying uncertainty should be provided to decision makers [3]. One way of doing this is to visualize uncertainty in a manner that is both intuitive and comprehensive. When designing an optimal visualization method for uncertainty, the varying goals, environment and types of information must be considered [4]. Many methods have been developed in past decade, but there is a little real world verification, that uncertainty visualization has been helpful [5].

In this poster, we introduce an ongoing project where uncertainty-aware drainage divides were calculated, visualized, and tested as background data for the decision-making process. Our objective was to study whether uncertainty information has an impact on decision making. Our study was part of a larger project in which the goal was to update the Finnish Drainage Basin System and Register in 2009-2013, taking into account the user requirements and INSPIRE specifications.

2 Materials and methods

The project involved three sets of spatial data: A digital elevation model, laser scanner data and hydrographic data. This data were used together with a topographic map of Finland, which supplied the background for the visualization experiments. Three firstly mentioned were input for a process to calculate uncertainty-aware drainage divides in our test area, the drainage basin of the Vantaa River, which locates in southern coast of Finland and covers an area of 1700 km^2.

The propagation of DEM errors for the drainage divide uncertainties was carried out using the Monte Carlo simulation method [6]. The uncertain drainage divide surface was generated by repeating the catchment delineation 400 times, each time with a different DEM. The values in the resulting surface represent the probability of each pixel to lie on the catchment boundary.

The uncertain catchment boundaries were visualized using seven methods, which differ from each other based on their color scheme, level of generalization (continuous/ categorized), and data model (raster/point). The visualization methods employed included:

A) A single hue mask,
B) A continuous color ramp, where the lightness of a single hue changes (light blue – dark blue),
C) A categorized color ramp, where the lightness of a single hue changes (light purple – dark purple),
D) A continuous color ramp between two hues (yellow – brown),
E) A continuous color ramp, where the lightness of a single hue changes (light purple – dark purple),
F) A continuous color ramp between three hues (blue – yellow – red), and
G) A graduated point symbol representation.

Methods A, C, and G represent generalized data, whereas methods B, D, E, and F show all of the details of the uncertainty surface. The surfaces were visualized by 40%

transparency on top of a topographic map of Finland at scale of 1:8000.

2.1 User survey

In order to study the impact of uncertainty visualization on decision making, we organized a user survey. The participants were the end users of the Finnish drainage basin dataset and all of them worked in governmental agencies. We invited a select number of persons to take part in an internet query.

The questionnaire had three sections: (1) background questions, (2) decision-making tasks, and (3) comparison of the visualizations.

For the decision-making tasks, we displayed a point on top of two different catchment boundaries overlaid on the base map. The first map showed the boundary without uncertainty information (fig. 1), while the second map displayed the uncertainty information (fig. 2). After seeing each of the images, the user was asked in which drainage area and how likely the point belongs to.

For comparing the visualizations, we showed the user seven different representation of the same surface and asked which visualization was (1) the most easy to read, (2) the most informative, (3) the most visually pleasing, and (4) the best choice for drainage divide analysis. We also asked users if they needed uncertainty information in their work and in which situations they found it useful.

3 Preliminary results

The preliminary results showed that providing uncertainty information did have an impact on decisions. Nearly 60% of the participants changed their answers after being provided uncertainty information and 76% found information about uncertainty useful in their work. The visualization comparison answers were scattered among the different methods, but regardless of the criteria the categorized color ramp and the continuous color ramp between yellow and brown received more votes than the other methods.

Figure 1: Sample decision-making task: The drainage divide is represented without uncertainty information and the user is asked in which drainage area (H or I) and how likely the point belongs to.

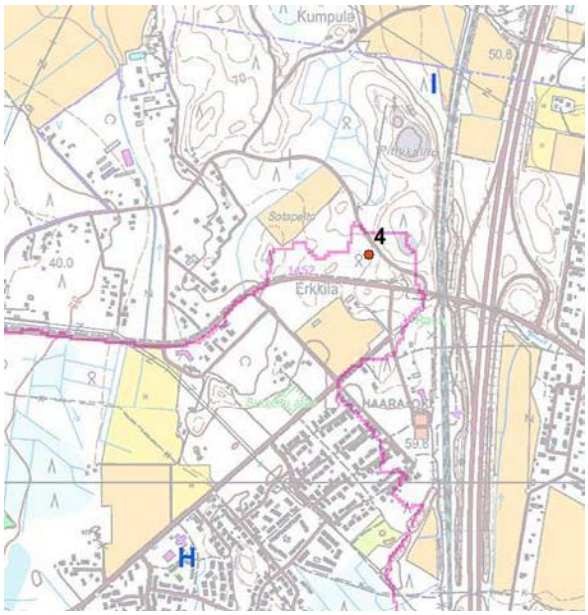
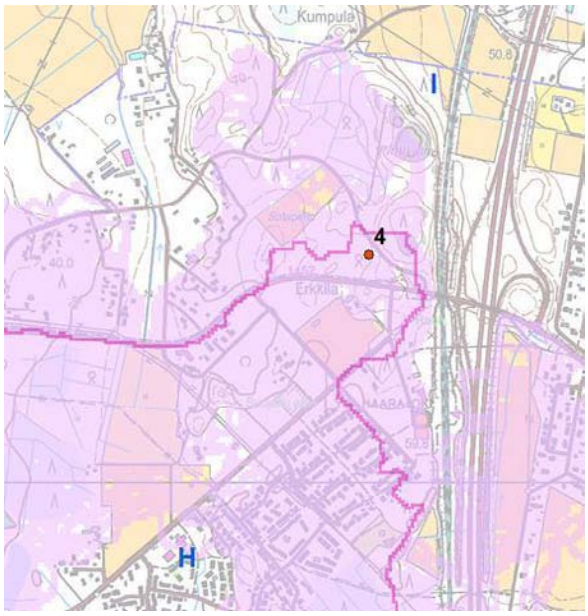


Figure 2: Sample decision-making task: The drainage divide is represented with uncertainty information and the user is asked in which drainage area (H or I) and how likely the point belongs to.



References

[1] Inspire-network and National Council for Geographic Information. Location combines – Finnish National Geographic Information Strategy 2010-2015, *Publication of the Ministry of Agriculture and Forestry*, 3/2010, pp. 28, 2010.

[2] S. Hope and G.J. Hunter. Testing the effects of positional uncertainty on spatial decision making. *International Journal of Geographical Information Science*, 21(6), pages 645-665, 2007.

[3] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *Conference on Simulation and Visualization*, pages 143–156, 2006.

[4] A.M. MacEachren. Visualizing uncertain information. Cartographic Perspectives, no. 13, pages 10–19, 1992

[5] A.M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan and E. Hetzler. Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science*, Vol.32, No.3, July 2005, pages 139-160, 2005.

[6] J. Oksanen. Digital elevation model error in terrain analysis. Ph.D. thesis. University of Helsinki, Faculty of Science, Department of Geography, 2006.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

SOS Server Deployment for Sharing Environmental Sensor Data Through the OTALEX-C Spatial Data Infrastructure

Pedro Vivas White
CNIG, C/ General Ibáñez de Ibero, 3.
28003
Madrid, Spain
pedro.vivas@cnig.es

Ignacio Brodin
PRODEVELOP, S.L,
Pza. Don Juan de Villarrasa,14,5. 46001
Valencia, Spain
ibrodin@prodevelop.es

Amelia del Rey
PRODEVELOP, S.L,
Pza. Don Juan de Villarrasa,14,5.46001
Valencia, Spain
adelrey@prodevelop.es

Jorge Sanz
PRODEVELOP, S.L,
Pza. Don Juan de Villarrasa,14,5. 46001
Valencia, Spain
jsanz@prodevelop.es

Abstract

OTALEX Spatial Data Infrastructure (SDI) project is funded by the INTERREG III European Programme and its main objective is to study and show the reality of the territory composed by the regions Alentejo in Portugal and Extremadura region in Spain. The current project phase is called OTALEX C and it is focused on showing the results of environmental studies through the OTALEX SDI. For that goal, it is necessary to catalogue, standardize, geoprocess and publish data from environmental sensors, as well as publish thematic contour and continuous maps derived from the interpolation of these sensors data. The first part of the project was the development of the processes to load a heterogeneous group of data from different types of sensors into a central repository using open source and self-developed Extract Transform and Load tools (ETL).

1 Introduction

The SDI OTALEX project is funded by the European Programme INTERREG III and its main objective is to study and show the reality of the territory, made up by regions of Alentejo in Portugal and Extremadura in Spain. Both regions are separated by country frontiers but share physical, environmental, social and economic characteristics.

The current phase of the project called OTALEX-C includes environmental sensor data publishing, as well as thematic maps publishing. These thematic maps are composed of isolines and continuous maps from the interpolation of environmental sensor data.

The project involved data transformation and load, standardization and integration of data from heterogeneous environmental sensors using the SOS-T protocol and thematic maps generation.

2 Evolution of the Project

Within the framework of OTALEX project, an automatic system was created to catalogue, standardize, geoprocess and publish data from heterogeneous environmental sensors, as well as publish thematic contour and continuous maps derived from the interpolation of the sensors data.

Figure 1: Current Geoportal of OTALEX SDI.

Source: OTALEX IDE web page: www.ideotalex.eu

2.1 Data Capture from Internal Data

A portable environmental station was acquired to measure in real time the following parameters:

- Wind's direction and speed
- Air pressure
- Relative humidity
- Temperature
- IR, UV, Beta and Gamma radiation

To access the data from the portable environmental station, some ETL processes were developed. These processes allowed to integrate the observations stored in the main Access/SQL database by the program GEONICA SUITE into the main repository of OTALEX-C project (Postgres/PostGIS). The ETL processes were developed with the open source GeoKettle tool, having direct access to the Access/SQL database and discovering the sensors for its registration into the SOS server.

2.2 Data Capture from External Data

A study to detect the environmental data sources that are useful in this project was carried out. The selected sources were:

- Extremadura Protection and Research Quality Network (REPICA). This network provides the following data: Sulfur Dioxide, Carbon Monoxide, Wind's direction and speed Wind, Temperature, Relative humidity, etc. The incorporation of these data to the SOS service was made by receiving daily data e-mails. This data was imported it into the main database of OTALEX by using ETL processes.

- Extremadura Irrigation Advice Network (REDAREX). This network provides the following data: Sun radiation, Relative humidity, Pluviometer, Wind's direction and speed and Temperature. In this case, as the data was not published through standards, a HTML scrapping process was carried out, in order to store these data into the central OTALEX database.

- Évora University. This station provides the following data: Solar radiation, Ultraviolet Solar radiation, atmospheric radiation, Temperature, Infrared radiation, etc. These data were imported into the main database of OTALEX by using ETL processes from an FTP server.

All these external data become the main kernel of data to build thematic maps with the environmental sensor measurements.

2.3 Central Database Definition

In order to store the sensors data, a scheme was deployed over a PostgreSQL/PostGIS 9.1. It was also created a scheme with different views for the simplification of the suitable information to be consumed by the thematic map server, as it will be explained ahead.

2.4 SOS Server Configuration and Implementation

In order to make possible the standardization and integration of data from environmental sensors a SOS server was integrated in the current architecture of OTALEX SDI, taking into account the standard SWE (Sensor Web Enablement).

A 52 North server was implemented, as well as the implementation of a SOS client inside the Geoportal to display and query current and past sensors observations.

All datasets from sensors were not directly loaded into the database but through the SOS-T protocol.

The SOS-T protocol allows to abstract the process of the schema of the database, allowing keep the process, even changing the SOS or the database software of the system.

The last server release 3.x that supports SOS 2.0 and SOS 1.0 was installed.

The implementation of a SOS client inside the Geoportal to display and query the repository of environmental data was made by using the SOS OpenLayers client. This SOS client makes possible to add SOS servers and point layers over the map.

2.5 Transformation and Load of Data into the Main Repository

To load the heterogeneous data from sensors into the central repository PostgreSQL/PostGIS a self developed ETL(Extract Transform and Load) process were designed by using the open source GeoKettle project.

On the other hand, other processes were developed for scrapping of data in HTML format.

The results of both processes were inserted into the database through HTTP requests to the SOS server.

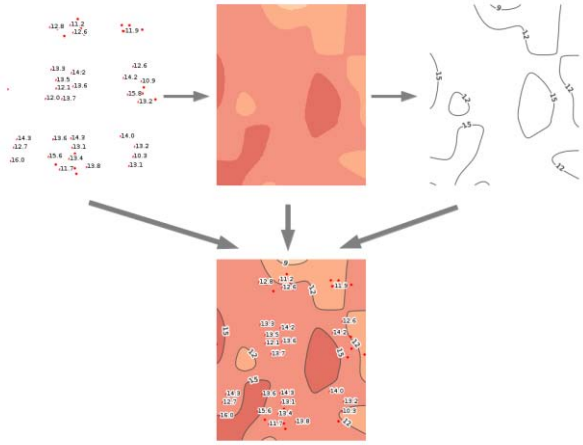
2.6 Thematic Maps Publishing

The OTALEX-C project uses GeoServer to publish cartography, through OGC services. Two new features of this server were used:

- SQL Views [1]: The traditional way to access database data is to configure layers against either tables or views. With this feature, layers can also be defined as SQL Views. These SQL Views allow to execute a custom SQL query on each request to the layer.
- Rendering Transformations [2]: They are invoked within SLD styles. Parameters may be supplied to control the appearance of the output. The rendered output for the layer is produced by applying the styling rules and symbolizers in the SLD to the result of transformation.
  There are three processes supported: Heat maps creation, Barnes Interpolation [3] and contour lines extraction from a raster.

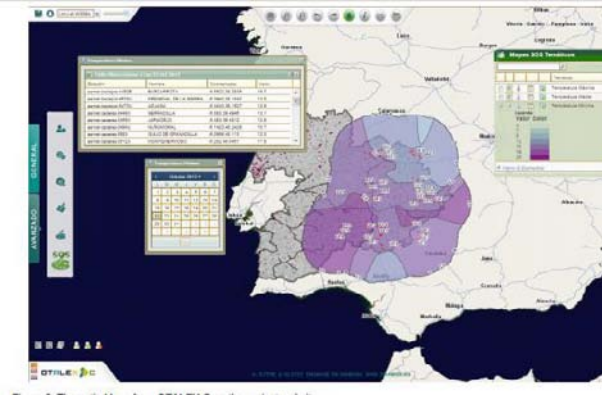
In OTALEX, these two features enabled to create maps of interpolation from a layer of points with the environmental data. A second style links the interpolation with the contour lines.

Figure 2: Styles Composition with rendered transformations



Source: OTALEX IDE website : www.ideotalex.eu

Figure 3: Thematic Maps in OTALEX-C

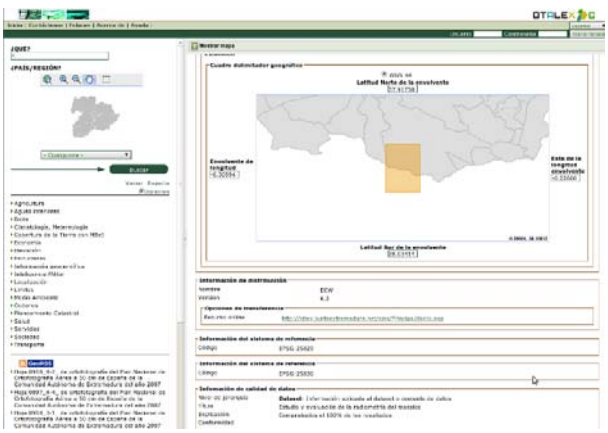


Source: OTALEX IDE website : www.ideotalex.eu

2.7 Metadata of SOS and WMS Services Creation

The SOS and the WMS services were included into the Metadata Catalogue of OTALEX. Thereby these services are available for its query through OTALEX SDI.
The metadata of these both services were created by using GeoNetwork and published by using the CSW protocol.

Figure 4: Metadata Catalogue in OTALEX



Source: http://www.ideotalex.eu/geonetwork/

3 Conclusions

Some of the main conclusions obtained along the development of this project were:

- The information from sensors should be published by using structured and consistent formats in order to have the possibility to integrate them and to share between different organizations. It's desirable to achieve the Open Data 4th level [4] for published sensor data.
- The use of SWE standards and protocols allows to integrate environmental information from heterogeneous data sensors, making possible the study of the reality of the territory

References

[1] Geoserver, Working by databases/SQLViews: http://docs.geoserver.org/stable/en/user/data/database/sqlview.html

[2] Geoserver, Rendering Transformations: http://docs.geoserver.org/stable/en/user/styling/sld-extensions/rendering-transform.html

[3] Barnes, Stanley L.(1964) A Technique for Maximizing Details in Numerical Weather Map Analysis. J. Appl. Meteor., 3, 396–409,1964.doi:http://dx.doi.org/10.1175/1520-0450(1964)003<0396:ATFMDI>2.0.CO;2

[4] Tim Berners Lee (2006), Linked Data: http://www.w3.org/DesignIssues/LinkedData.html

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Design of the Data Transformation Architecture for the INSPIRE Data Model Browser

A. Belussi, S. Migliorini
Department of Computer Science,
University of Verona – Italy
alberto.belussi@univr.it

P. Cipriano
EC-JRC
Ispra – Italy
pg.cipriano@gmail.com

M. Negri, G. Pelagatti
Dep. of Electronics, Information and
Bioengineering, Politecnico of Milan – Italy
giuseppe.pelagatti@polimi.it

Abstract

The INSPIRE directive requires that inside a Spatial Data Infrastructure (SDI) data are provided using a model compliant with the INSPIRE Data Model (IDM). Therefore, one of the main issues during the implementation of an SDI is the transformation of existing source databases in the way defined by the IDM. In literature, many aspects of the INSPIRE transformation problem have been studied and classified, this paper deals with the definition of the transformation architecture and the schema transformation levels.

1 Introduction

One of the main problems facing the implementation of the European Spatial Data Infrastructure foreseen by the INSPIRE directive is the transformation of data stored in existing source databases into the model defined by the INSPIRE data specifications. These specifications are formally defined by a set of application schemas, called *INSPIRE schemas*; which are based on a model, called *INSPIRE Data Model* (IDM). The IDM definition is mostly contained in the *INSPIRE generic conceptual model* [1], which in turn refers to many EN ISO 19100 Standards [2]. The aforementioned transformation may be called *INSPIRE transformation*.

The *INSPIRE Data Model Browser* (IDMB) is a tool that allows one to represent an INSPIRE schema in a form which is simpler to read by non-UML-experts. It is derived from an operational tool, the GeoUML catalogue, which has been funded by Italian local governments, produced by Politecnico of Milan with CISIS (Centro Interregionale per i Sistemi Informatici geografici e Statistici), and used for the specification of the Italian National Core content, and by Italian Regions and other public bodies for particular content specification. IDMB will be distributed freely since 2014, also through the European Commission Joinup Platform (https://joinup.ec.europa.eu/community/are3na/home).

This paper analyzes the data transformation problem in order to derive the requirements for extending IDMB with data transformation capabilities. This paper assumes without loss of generality that the source databases are Spatial SQL databases.

The INSPIRE transformation approaches have been extensively studied and analysed [3-6] and some tools that perform to some extent this kind of transformation are available. However, a completely satisfactory solution has not been found yet, and data producers facing with the difficult problem of choosing both methodology and tools to satisfy the INSPIRE requirements.

2 Main Transformation Approaches

Two basic types of *schema transformations* from a source schema, defined with a Source Data Model (SDM), into a target schema, defined in a Target Data Model (TDM), can be classified:

- *Content transformation*: the source and target models are the same (SDM=TDM).
- *Model transformation*: the source and target models are different (SDM≠TDM), but no content transformation is performed: the performed schema transformations do not depend on the schemas, but only on the models.

A schema transformation which is neither a content nor a model transformation is called *mixed transformation*. In general, an INSPIRE transformation requires to perform a mixed transformation where SDM=SQL, TDM=GML, source content= "Content of the source database", and target content= "Content of INSPIRE specification".

Since in a mixed transformation the most difficult part is due to the content aspects, it is useful to perform this part in a common model situation; i.e., to perform it as a content transformation such that SDM=TDM. Therefore, it is convenient to decompose a mixed transformation into a content transformation and model transformation. Two basic approaches that can be define by considering the possible orders of these two transformations:

A. The *model-content approach*: first apply a model transformation of the source database to GML and then apply a content transformation using GML.
B. The *content-model approach*: first apply a content transformation from the source database to a new database, often called the INSPIRE database, using SQL, and then apply a model transformation from the INSPIRE database to GML. This is the approach considered in this paper and its details are reported in Table 1.

The fundamental differences between the two approaches are:

- The environment where the data transformation is executed: GML in the first approach and an SQL database in the second one.
- The model transformation (WFS configuration) must be performed for each different source database in the first approach, while it is defined only once for the INSPIRE database in the second one.

Table 1: The content-model transformation approach.

	Level	SDM		TDM
Content transformation	model	SQL		SQL
	schema	SQL/DDL	SQL transform.	SQL/DDL
	instance	Source DB	derived SQL/DML scripts	InspireDB
Model transf.	model	SQL		GML
	schema	SQL/DDL	WFS conf. (only once)	GML XSD (Inspire)
	instance	Inspire DB	WFS	GML data

3 Conclusion

In the *content-model approach*, the implementation of the target classes as relational tables allows one to perform the transformations by SQL queries. The use of an SQL transformation allows one to guarantee the "stability" of new objects and identifiers. Moreover, in the SQL queries spatial indices can be used in order to improve the performance of transformations based on geometric properties. In order to automate the transformation process, it is possible to define some template, which can be used to automatically generate all transformations sharing the same basic structure. The definition of such templates has highlighted the necessity to save some partial results and perform the transformation in several steps; hence performing it in a database environment is convenient (if not necessary).

References

[1] European Commission. (2013). *INSPIRE Generic Conceptual Model*. [Online]. Available: http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3.4rc3.pdf [Accessed: 27-Nov-2013]

[2] European Committee for Standardization. (2013). CEN/TC 287 published standards. [Online]. Available: http://www.cen.eu/cen/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/Pages/Standards.aspx?param=6268&title=CEN/TC%20287 [Accessed: 27-Nov-2013]

[3] F.Arntsen and M. Borrebaek. (2013). "From production data base to INSPIRE data using WFS: potential methods" presented at the INSPIRE KEN workshop. [Online]. Available: http://www.eurogeographics.org/content/inspire-ken-eurosdr-workshop [Accessed: 27-Nov-2013]

[4] M.L. Vautier, "Schema Transformation Concepts", ibidem

[5] S. Balley, S.Mustière and N. Abadie, "Involving Semantics in schema transformations", ibidem

[6] T. Reitz, "Classifying Schema Transformation Approaches and Tools", ibidem

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Error propagation in a fuzzy logic spatial multi-criteria evaluation

Lisa Bingham
University of Stavanger
Department of
Petroleum Engineering
Stavanger, Norway
lisa.bingham@uis.no

Derek Karssenberg
Utrecht University
Department of Physical
Geography
Utrecht, the Netherlands
d.karssenberg@uu.nl

Abstract

Quantifying errors in results of spatial multi-criteria evaluation (MCE) techniques is essential to improve the credibility of MCE in planning and decision-making. We present an error propagation procedure using Monte Carlo simulation for fuzzy logic MCE applied in a case study of petroleum exploration which covers northern South America. The fuzzy logic MCE combines data sets to evaluate the favourability of petroleum exploration in a geographic region. Each input data set has associated error models and estimated uncertainty. Two sources of error are investigated: boundary and fuzzy membership. 2000iterations of the model were run. The resulting mean of the 2000 samples and a series of confidence interval maps were analysed. It is concluded that the combination of the MCE analysis and error propagation modelling will support decisions for petroleum exploration.
Keywords: error propagation, fuzzy logic, multi-criteria evaluation, petroleum.

1 Introduction

Multi-Criteria Evaluation (MCE) is a subset of multidimensional decision and evaluation models that essentially are tools to evaluate the trade-offs between alternatives with different impacts [3]. The goal of MCE is to evaluate the outcome of combining different criteria to fulfil one or more objectives that may possibly be conflicting [3]. It is important to understand that all types of MCE are subjective and different methods may give different results (e.g. [5]).

Bingham et al. [1] proposed a spatial MCE method based on a series of user-defined inputs and criteria that can be used to show geographic areas that may be of interest for further investigation for petroleum exploration (Figure 1). By using the results of the analysis, the geologist can support his reasons for requiring more or less investigation in a geographic region for future exploration.

The purpose of the model is to produce maps for petroleum exploration; high-scoring areas are more favorable for petroleum exploration based on the criteria (the input data sets; the top level of boxes shown in Figure 1). The input data sets are assigned fuzzy membership values [0,1] where 0 is unfavorable and 1 is favorable. The MCE uses fuzzy logic operators (FLOs; AND, OR, GAMMA, ALGEBRAIC SUM) to combine the input data sets. The operators work on a cell-by-cell basis (i.e. raster algebra); thus all input data sets must be in the same coordinate reference system (projection), have the same cell size, and cover the same spatial extent [2].

Following the model framework (Figure 1), the favorability of petroleum exploration can be modeled either by specific age intervals which would be appropriate for determining drilling targets or by examining all of the available data (i.e. non-age-specific) which would be appropriate for a general overview of petroleum exploration favorability.

All models, including MCE, have some amount of uncertainty related to their input, parameters, and results, which when unknown decrease the reliability of decisions based on the output [6, 7]. The fuzzy logic MCE proposed has some uncertainty associated with the input data sets. This paper applies error propagation using Monte Carlo simulation to a fuzzy logic multi-criteria evaluation. The original work [1] acknowledged that there is some uncertainty with the results but it was not quantified or evaluated thoroughly. This work aims to rectify this lack of information. It specifically investigates:

- What kind of error models should be applied given the data?
- In the case study, how do the confidence intervals and mean compare to the original result?
- How reliable are the results from the MCE?

The sources of uncertainty are neither gross (i.e. due to negligence or carelessness) nor systematic (i.e. having a functional relationship), but random [6]. The error of the input data sets can be estimated; thus, error propagation is a suitable method for investigating the estimated error effects on the final output [6]. Uncertainty propagation modelling is important in order to assess how model input errors propagate to the model results, in order to quantify the uncertainty in the model results [5, 8]. Any uncertainty propagation modelling is subjective based on the parameters of the uncertainty modelling.

Monte Carlo simulation is a technique, which calculates the model repeatedly using different input values based on an error model [4]. By interpreting the culminated results (e.g. 95% confidence interval, median) of hundreds or thousands of samples, the modeller can determine if the model results are reliable.

2 Approach

Three main sources of uncertainty in a model are parameters, inputs, and model structure [7]. Uncertainty related to inputs is investigated in this study; parameters and

model structure are not investigated. Input data uncertainty is common in a GIS environment where the user relies on many different sources for data, especially if data is older or the original source is unknown. Border classification and fuzzy membership assignment are the only sources of uncertainty that are investigated in this paper.

The workflow of the uncertainty error modelling follows a series of steps to create realizations for the investigated uncertainties. After a number of realizations are completed, the composite probabilities, average, and variance are calculated.

3 Case study

The study area focused on northern South America from Colombia in the west to Guyana in the east; the data was compiled and converted to PCRaster format following general conventions. For more details on the data compilation the reader is referred to Bingham et al. [1]. The authors assume that the attribute classifications of the data are correct and do not contain gross or systematic errors. The uncertainty for all input data sets is estimated to be 10%, except in the case of geochemical and seeps data which were estimated to be 30%.

The size of the study area covers more than 2.5 x 10^6 km^2. Data were compiled from sources ranging in scale from 1:100,000 to 1:44,000,000 and converted to raster format using a customized Albers projection and a cell length of 1000 m.

The Monte Carlo simulation ran 2000 samples for a non-age-specific model. Each simulation took approximately 15 hours on PCRaster v. 3.0 with python v. 2.6 on a CentOS 5, 64-bit, 8 Xeon 2.66 GHz processors, and 32 GB RAM.

3.1 Results

The Monte Carlo simulation focused on the non-age-specific favorability model. This model included all of the data meeting the criteria regardless of age. The original model result is shown in Figure 2 for comparison. Figure 3 shows the output map of the average from 2000 samples. The original model result (Figure 2) is very similar to the average; both maps have approximately the same maximum range (0.78 vs 0.79) and reflect the same pattern for high versus low favorability areas. The average of the 2000 samples shows more and larger areas of high favorability. Figure 4 shows the 2000 samples at a 95% confidence interval with a threshold value of 0.7; in the original model results, this value was arbitrarily chosen as an indicator of high favorability where additional exploration may be justified. This combination of confidence interval and threshold value results in areas that are certain to be below a favorability value of 0.70, areas that are certain to be above this value, or areas for which it is uncertain whether the favorability will be above or below 0.70. In this map, areas where it is certain the favorability value is above 0.70 are indiscernible. However, by decreasing the threshold value to 0.50, some areas are now certain to be above the threshold value (Figure 5). If the exploration geologist finds it appropriate, the confidence interval can be increased to 50% and more areas will lie above the threshold value (Figures 6 and 7).

4 Conclusions

The Monte Carlo simulation investigating error propagation of fuzzy logic multi-criteria evaluation shows that the model and its inputs are reasonably reliable; however, the data itself may be improved to reduce the amount of uncertainty. The confidence interval and threshold value must be chosen with care and communicated to the end user. With improved data, the multi-criteria evaluation in conjunction with the Monte Carlo simulation can provide a useful tool for decision-making in petroleum exploration.

References

[1] Bingham, L., R. Zurita-Milla, and A. Escalona, GIS-based Fuzzy Logic for Petroleum Exploration. *AAPG Bulletin*, 96:2121-2142, 2012.

[2] Bonham-Carter, G. F., *Geographic Information Systems for Geoscientists Modelling with GIS: Computer Methods in the Geosciences Volume 13*: Oxford, Pergamon, 1994.

[3] Carter, S., Site search and multicriteria evaluation. *Planning Outlook*, 34:27-36, 1991.

[4] Heuvelink, G. B. M., *Error propagation in environmental modelling with GIS*. Taylor and Francis, London, 1998.

[5] Heywood, I., J. Oliver, and S. Tomlinson, Building an exploratory multi-criteria modelling environment for spatial decision support. *Innovations in GIS*, 2:127-136, 1995.

[6] Thapa, K., and J. Bossler, Accuracy of Spatial Data Used in Geographic Information Systems. *Photogrammetric Engineering and Remote Sensing*, 58:835-41, 1992.

[7] Verstegen, J. A., D. Karssenberg, F. van der Hilst, and A. Faaij, Spatio-temporal uncertainty in Spatial Decision Support Systems: A case study of changing land availability for bioenergy crops in Mozambique. *Computers, Environment and Urban Systems*, 36:30-42, 2012.

Figure 1: Workflow of the model. Data sets are combined starting at the top level and moving subsequently down.

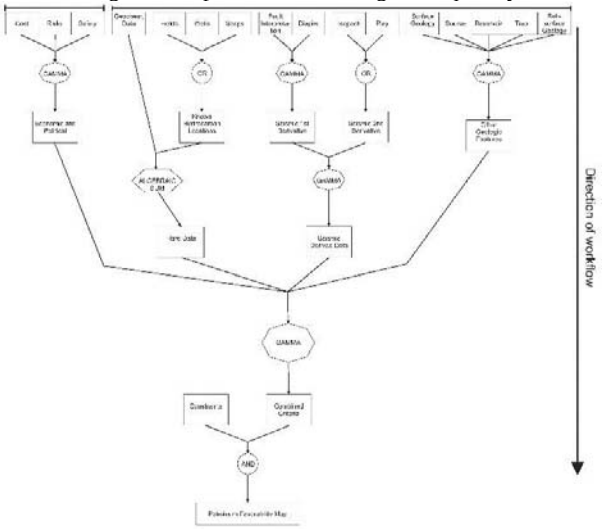


Figure 2: Original result showing petroleum exploration favorability [1]. Darker grey reflects higher favorability..

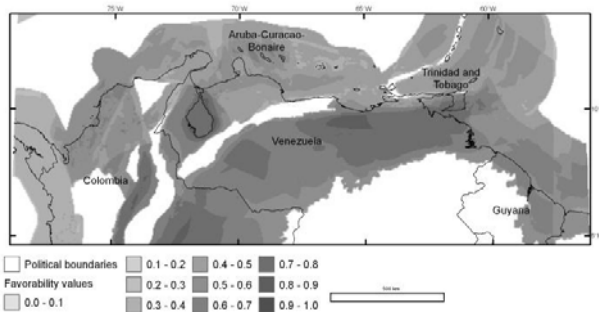


Figure 3: Average result of 2000 samples showing petroleum exploration favorability. Darker grey reflects higher favorability.

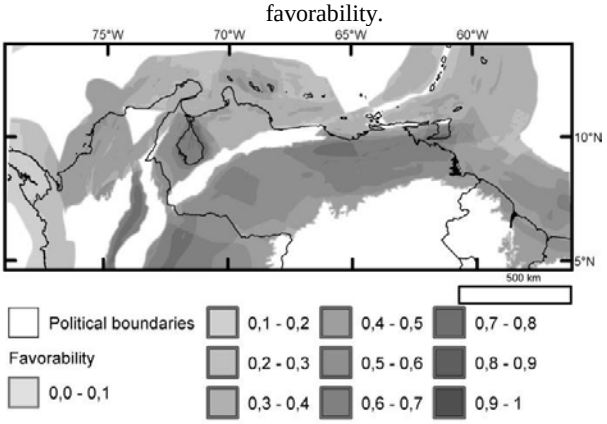


Figure 4: 95% confidence interval with threshold 0.70.

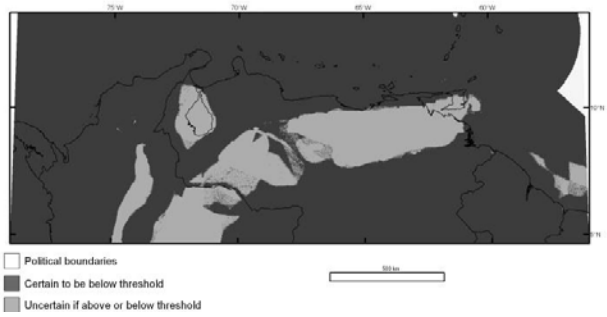


Figure 5: 95% confidence interval with threshold 0.50.

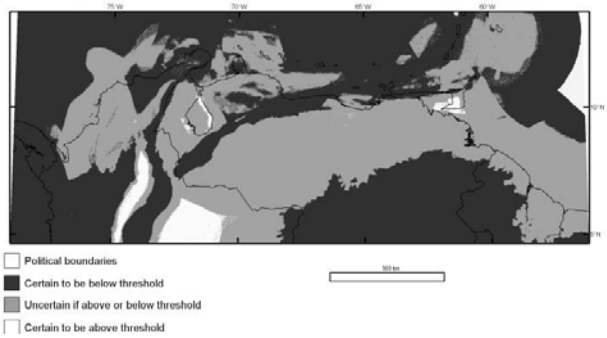


Figure 6: 50% confidence interval with threshold 0.70.

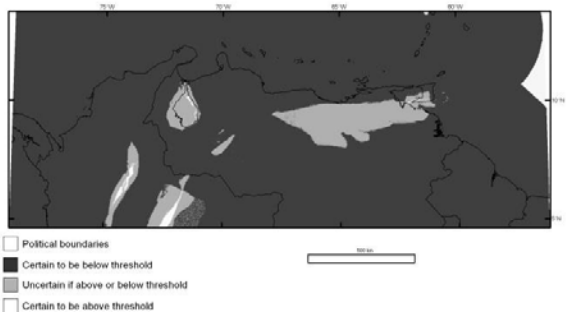
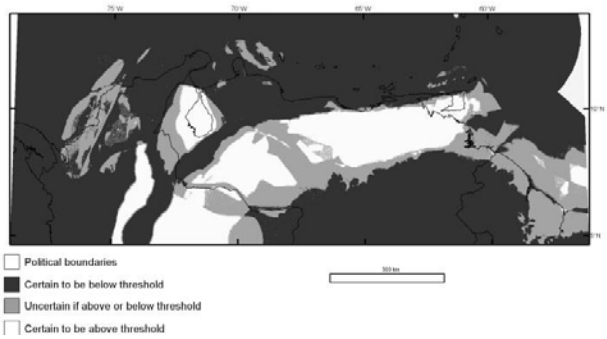


Figure 7: 50% confidence interval with threshold 0.50.



Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

ELF GeoLocator Service

Pekka Latvala
Finnish Geodetic Institute
Masala, Finland
pekka.latvala@fgi.fi

Lassi Lehto
Finnish Geodetic Institute
Masala, Finland
lassi.lehto@fgi.fi

Jaakko Kähkönen
Finnish Geodetic Institute
Masala, Finland
jaakko.kahkonen@fgi.fi

Abstract

This paper describes the implementation of a gazetteer service, GeoLocator, developed in the project 'European Location Framework' (ELF). The GeoLocator service contains data from the INSPIRE/ELF themes Geographical Names, Administrative Units and Addresses. The functionalities of the service include geocoding, administrative unit-limited geocoding, fuzzy geocoding, reverse geocoding and administrative unit-limited reverse geocoding.

Keywords: Gazetteer Service, geocoding, reverse geocoding, WFS-G.

1 Introduction

The ongoing implementation of the INSPIRE directive is resulting in the creation of harmonized European-wide spatial data that covers multiple data themes and enables the creation of spatial Web services that provide functionalities for the whole area of Europe.

This work describes the process of implementing the ELF GeoLocator service that is a gazetteer service that contains multilingual and authoritative European spatial data.

The implementation of gazetteer services have been standardized by the Open Geospatial Consortium (OGC) in the gazetteer service application profile of the Web Feature Service interface (WFS-G AP) best practice paper in 2012 [1].

The ELF GeoLocator service is based on earlier EuroGeoNames (EGN) service, originally created in 2006-2009 [2] and renewed in 2012 [3]. The main objectives in the project are (1) to add more geographical names (GN) data into the service and to import new data from the themes 'Addresses' (AD) and 'Administrative Units' (AU) and (2) to create new service functionalities by enhancing the service's geocoding capabilities and by adding new reverse geocoding functionalities.

2 Related Work

Digital gazetteers are an important research topic in geoinformatics. The core elements of digital gazetteers have been studied by Hill [4]. Some examples of gazetteer services are a meta-gazetteer service [5] that integrates data from multiple gazetteer services and supports geocoding and reverse geocoding and a geoXwalk [6] gazetteer that parses place names from various documents.

There are currently many gazetteer services available on the Internet. Manguinhas et al. [7] have collected a list of many of these services.

3 ELF GeoLocator Service

3.1 Service Architecture

The ELF GeoLocator Service is based on a centralized architecture that contains a PostgreSQL/PostGIS service database that stores all the collected data (Figure 1). The database contains the contents of the EGN database and AD, AU and GN data that have been collected from the national INSPIRE/ELF download services.

The main service output is the WFS-G AP encoded output that is created with the deegree WFS application. On top of it there is a custom front-end WFS module, developed in the EGN and ELF projects. It is a Java Servlet module that handles the execution of the WFS query process. The front-end WFS adds the support for custom service operations and the support for custom LANGUAGE parameter that can be used with the WFS-G output for requesting the location type information in a specified language.

3.2 Service operations

The service supports the WFS operations *GetCapabilities*, *DescribeFeatureType* and *GetFeature*. The main geocoding functionality is available through the *GetFeature* operation by using filters with the queries according to the OGC's filter encoding language.

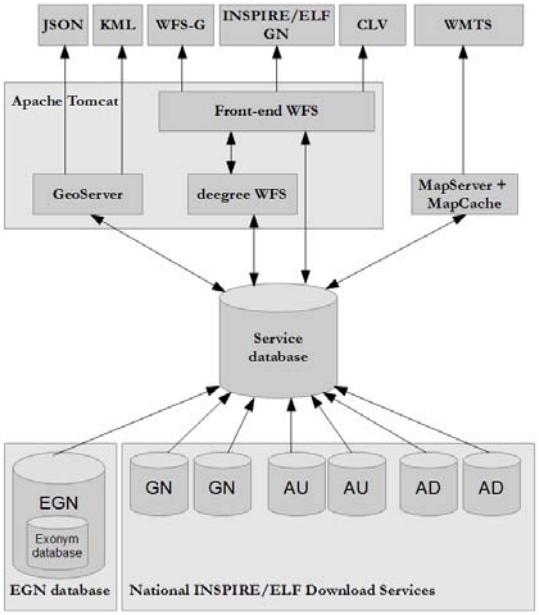
The service supports also three custom operations *GetFeatureInAu, FuzzyNameSearch* and *ReverseGeocode.*

The *GetFeatureInAu* operation limits the geocoding inside a specific administrative unit. It is useful for finding results when there are multiple features that have the same name.

The *FuzzyNameSearch* operation executes name searches that can find features from a slightly misspelled input. It is useful when the user makes a typing error or when the queried name contains diacritics or special characters.

The *ReverseGeocode* operation contains two modes: (1) normal mode where the operation returns the feature nearest to the given coordinate point. (2) AU–limited mode where the operation returns the AU-based feature from the most detailed AU level that contains the given coordinate point.

Figure 1: The architecture of the ELF GeoLocator service



4 Discussion

Currently the ELF project is ongoing and the total amount of countries that are able to provide the data for the ELF GeoLocator service is unknown. The eventual aim is the full coverage of the EuroGeographics member countries.

In future the ELF GeoLocator service could be expanded with new operations and new data from new themes. One potential data theme is the INSPIRE theme Cadastral Parcels (CP).

5 Conclusions

The ELF GeoLocator service provides functionalities that are fundamental in the spatial data infrastructure. It enlarges the data contents of the EGN service by increasing the amount of its GN data and by importing data from new themes: AD and AU. The service functionalities are expanded with administrative unit-limited geocoding, fuzzy geocoding, reverse geocoding and administrative unit-limited reverse geocoding. In future the service will be further developed by uploading content from new countries via the INSPIRE/ELF-compliant Download Services provided by the participating NMCAs.

References

[1] J. Harrison and P.A. Vretanos, editors, Gazetteer Service – Application Profile of the Web Feature Service Best Practice, 2012. Available at: https://portal.opengeospatial.org/files/?artifact_id=46964

[2] P. G. Zaccheddu, D. Overton, EuroGeoNames (EGN) – Implementing a sustainable European gazetteer service, UNGEGN Working paper No. 38, 2011. Available at: http://unstats.un.org/unsd/geoinfo/UNGEGN/docs/26th-gegn-docs/WP/WP38_EGN_item%209_UNGEGN26.pdf

[3] P. Latvala, L. Lehto and J. Kähkönen, The Renewed Implementation of the EuroGeoNames Central Service, *16th Agile Conference on Geographic Information Science*, 14-17 May, 2013, Leuven, Belgium. Available at: http://www.agile-online.org/Conference_Paper/CDs/agile_2013/Posters/P_Latvala.pdf

[4] Linda. L. Hill, Core Elements of Digital Gazetteers: Placenames, Categories and Footprints. In *Proceedings of the 4th European Conference, ECDL 2000*, Lisbon, Portugal, September 18-20, 2000.

[5] P. D. Smart, C. B. Jones and F. A. Twaroch, Multi-source Toponym Data Integration and Mediation for a Meta-Gazetteer Service, In *Proceedings of the 6th International Conference, GIScience*, 2010 Zurich, Switzerland, September 14-17, 2010.

[6] J. Reid, geoXwalk – A Gazetteer Server and Service for UK Academia, In *Proceedings of the 7th European Conference, ECDL 2003*, Trondheim, Norway, August 17-22, 2003.

[7] H. Manguinhas, B. Martins, J. Borbinha and W. Siabato, The DIGMAP geo-temporal web gazetteer service. In *Proceedings of Third International Workshop Digital Approaches to Cartographic Heritage*, Barcelona, Spain, June 26-27, 2008.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

ENHANCING THE ROLE OF CITIZEN SENSORS IN MAPPING: COST ACTION TD1202

Giles Foody
School of Geography
University of Nottingham
Nottingham, UK
giles.foody@nottingham.ac.uk

Steffen Fritz, Linda See
IIASA, Laxenburg
Austria
fritz@iiasa.ac.at
see@iiasa.ac.at

Norman Kerle
Faculty of Geoinformation Sci & Earth Observation
University of Twente The Netherlands
n.kerle@utwente.nl

Glen Hart
Ordnance Survey
Southampton
SO16 0AS, UK
glen.hart@ordn--ancesurvey.co.uk

Cidalia Fonte
Mathematics Dept
University of Coimbra
NESC Coimbra
Coimbra
Portugal
cfonte@mat.uc.pt

Abstract

This article introduces a strategic initiative, COST Action TD1202, focused on the role of citizen sensors in mapping. It outlines the Action's scope, aims and current status. In particular, the article outlines the potential of citizen science in mapping activities and indicates the scope of current work undertaken by the Action's four working groups. It is stressed that the Action is at an early stage and that it is open to new members.
Keywords: Mapping, citizen sensors, volunteered geographic information, COST Action.

1 Introduction

Over the last decade, citizen science has grown considerably and is especially evident in relation to geographical information with the rise of the citizen sensor [1]. This paper outlines a strategic initiative funded by the European Union to help enhance the role of citizen sensing in mapping.

2 COST Action TD1202

Cooperation in Science and Technology (COST) Actions are a European framework to support research on topics of global relevance. This paper introduces Action TD1202 on 'Mapping and the Citizen Sensor'.

Accurate and timely maps are a fundamental resource but their production in a changing world is a major scientific and practical grand challenge. Citizen sensing has the potential to radically change mapping. The quality of citizen sensor data, however, is variable and activity is often relatively uncoordinated.

This Action will evaluate the utility of citizen sensors in mapping and seek to encourage good practices while not constraining the activity of citizen sensors who often are unpaid volunteers.

2.1 Background

Mapping has benefited from recent advances in geoinformation technologies) including citizen sensors fostered by the proliferation of inexpensive and mobile location-aware devices able to provide supporting information (e.g. volunteered geographic information (VGI)) and a general increase in geo-literacy. But the full potential of citizen sensing is unrealized, especially because VGI has quality concerns, notably as sources range from naïve citizens to authoritative, but imperfect, agencies.

The Action aims to better understand the underlying motivation of contributors and to give recommendations on the incentives needed for a VGI project to succeed. Additionally, by bringing together an international team it encourages mobility and knowledge transfer. It began in November 2012 and is scheduled to run to November 2016. It involves 32 countries and welcomes new members (prospective members should contact Giles Foody).

Its central goal is to enhance the role of citizen sensors in mapping. It will review the current status of citizen sensors in mapping, evaluate the strengths and limitations of VGI for key tasks and add value to VGI by indicating its quality and steering activity in constructive ways.

2.2. Focus and Organisation

The Action's four working groups focus on overlapping topics.

WG1 focuses on acquiring and managing VGI. It aims to provide an understanding of current practices involving the acquisition, description, storage and distribution of VGI arising from citizen sensors. It has examined the array of terminology related to VGI and summarised the definitions and inter-relationships between the terms to provide clarity. It has also systematically evaluated VGI websites and mobile apps to characterise issues such as: the nature of data sources, the expertise and training of citizen sensors, mechanisms to make VGI available, meta-data provision and if quality control activity. It is planned to run data collection campaigns and to tap into the Action's network, with the data to be available to Action members for further analysis.

WG2 is focused on understanding and influencing contributors. It aims to develop an understanding of citizen sensors and their motivations. It is reviewing the motivation of volunteer types, and how that knowledge can be used to mobilize such groups, and what incentives or rewards are needed to make them contribute optimally and on a sustained basis. It also assesses how VGI campaigns of different types and size are best coordinated, how volunteers are best instructed and trained if needed, and how their contributions can best be assessed and validated. The insights gained should help promote active citizen sensing to meet specific needs.

WG3 addresses citizen sensing in map production. It aims to define the needs of the map producing community, identify the sensitivity and tolerance of mapping methods to different types of error and uncertainty in VGI and assess the potential role of current VGI efforts as well as of active citizen sensing. A survey of key map producers has been undertaken.

WG4 focuses on citizen sensing in map validation. It aims o characterize VGI quality assessment practices, identify the sensitivity and tolerance of different types of applications to different types of error and uncertainty in VGI and assess the potential role of current VGI efforts as well as active sensing. It is reviewing the methodologies used to assess the quality of VGI in several aspects, namely related to the contributor's reliability and traditional aspects of data quality (e.g. positional and thematic accuracy, completeness, currency and logical consistency). This WG will also contribute to the identification of best practices for use of VGI in validation.

3. Conclusions

Citizen sensors have considerable potential to aid mapping related applications. COST Action TD1202 seeks to enhance the role of citizen sensors in mapping. Work to-date has focused on reviewing the literature and it is open to new members able to contribute constructively.

Acknowledgments

We are grateful for the financial support provided by the EU for COST Action TD1202 'Mapping and the Citizen Sensor'.

References

[1] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *Geojournal* 69: 211-221. 2007.

[2] Committee on Strategic Directions for the Geographical Sciences in the Next Decade, *Understanding the Changing Planet: Strategic Directions for the Geographical Sciences*, National Academies Press, USA, 2010.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

IDE-OTALEX C. The big challenge of the first Crossborder SDI between Spain and Portugal

Teresa Batista
ICAAM-University of Évora; CIMAC
7000-673 Évora, Portugal.
tbatista@cimac.pt

Carmen Caballero
Gobierno de Extremadura
06400 Mérida, España
carmen.caballeroc@gobex.es

Fernando Ceballos
Gobierno de Extremadura
06400 Mérida, España
fernando.ceballos@gobex.es

Cristina Carriço
CIMAC
7000-671 Évora, Portugal
cristina.carrico@cimac.pt

Pedro Vivas
CNIG
28003 Madrid, España
pedro.vivas@cnig.es

José Cabezas
University of Extremadura
06071 Badajoz, España
jocafer@unex.es

Luis Ferdández
University of Extremadura
06071 Badajoz, España
luferpo@unex.es

Carlos Pinto-Gomes
ICAAM-University of Évora
7002-554 Évora, Portugal.
cpgomes@uevora.pt

Abstract

The SDI implementation is an average difficult work. There should be an understanding between political and scientific interests, technological advances and it is also quite recommended to meet the needs of citizens.

A cross-border SDI implementation, where three levels of administration belonging to two countries must be considered, may seem impossible to do, but it is not only a possible task, but also an enriching and useful task to study the reality of the territory and its sustainable development.

IDE OTALEXC is the first crossborder spatial data infrastructure characterized for being a distributed, decentralized, modular and collaborative system, based on standards OGC (Open Geospatial Consortium), W3C (World Wide Web Consortium), ISO (International Organization for Standardization) and open source technology.

1 Introdution

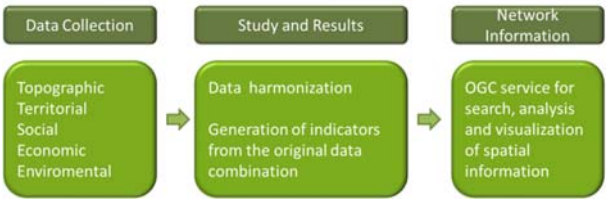
The Alentejo (Portugal), Extremadura (Spain) and recently Centro (Portugal) regions, are working along 16 years, several projects with the co-finance of the crossborder cooperation programs for Spain-Portugal of the ERDF (European Regional Development Fund) on study the reality and sustainable development of territory.

Three regions have similar characteristics; sparsely populated regions (less than 37 inhabitants per km2), whose main economic activities are agriculture and services and important environmental areas with several nature conservation sites and protected areas (Natura2000 sites, Birds Protection Areas and National Parks).

Figure 1: OTALEX C area.

IDE-OTALEX C has developed an Indicator System— SI-OTALEX C, to identify, measure, monitor and evaluate human pressures and its dynamics in the region. The established set of indicators has a common and standard structure designed by a multidisciplinary team with experts from both countries. Its main objective is to evaluate the transformations of the territory and help to solve common problems of the territory and their populations.

Figure 2: Basic schema of IDE-OTALEX C workflow.

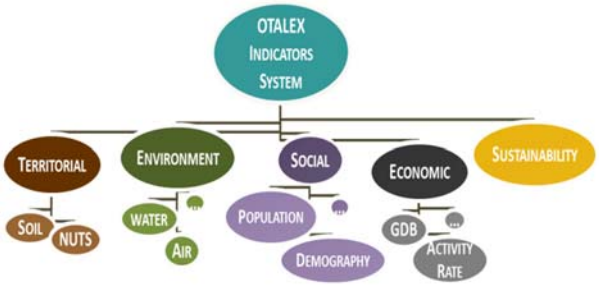


2 Data and Monitoring

Following the guidelines of the EU-SDS (Sustainable Development Strategy of the European Union), the national Development Strategies for Portugal and Spain, SI-OTALEX C was framed by the conceptual model PSR (Pressure-State-Response), adopted from Ref. [1]. Although, we know that

choices always involves disregarding something [1], since there is no universal set established, and there is such a high and diverse number of indicators, the core indicators for SI-OTALEXC was built with those that best fit the project objectives, had more relevance and representativeness in the area, and also easily available and measurable. Furthermore they should be simple, easy to read and update.

Figure 3: Basic OTALEXC Indicator System (SI-OTALEX C).



SI-OTALEX C has over sixty indicators grouped into twenty two themes, in turn, the themes, are grouped into five vectors as shown in the figure 3: Territory, Environment, Social, Economy and Sustainability

Table 1: SI-OTALEX C main structure.

VECTOR	THEME
01. Territory	01. Climate
	02. Geology and Geomorphology
	03. Hydrography
	04. Soil
	05. Administrative structure
02. Environment	01. Air
	02. Water
	03. Waste
	04. Pollution Sources
	05. Land use
	06. Environmental performances and Urban spaces
	07. Noise
	08. Energy
	09. Nature conservation
	10. Landscape
	11. Soil protection
03. Social	01. Population
	02. Demographic structure
	03. Equipment's and Services network
04. Economy	01. Economic activities
	01. Territorial matrix
05. Sustainability	02. Sustainable transport

3 IDE-OTALEX C

Spatial data infrastructures are, in general, for their characteristics, the best technological tool to publish sustainability data in the web. They can synthesize, calculate and analyse spatial data through interoperable web based services. SDIs are essential to manage natural resources, economic development and environment protection in a way to monitor the changes of the territory.

IDE-OTALEXC is the crossborder spatial data infrastructure of Alentejo, Extremadura and Centro. It was implemented in 2007 to share official geographic information between these regions. This has been the most effective way to have a distributed and flexible observatory for sustainable development and environment protection in this rural and low populated regions [3]. It also contributes to territorial cohesion, one of the tree main pillars of European Cohesion Policy.

Figure 4: Home page of website IDE-OTALEX C www.ideotalex.eu

IDE-OTALEX C is a distributed, decentralized, modular and collaborative system, based on standards (OGC, W3C, ISO) and open source technology, developed to guarantee interoperability between the different GIS provided by each project partner and OTALEX C itself.

Sam of the main results is:

- Data harmonization (cartography and indicators) on both sides of the border (fig. 5).
- Development of Analysis Tools through OGC standard WPS (Web Processing Services).
- Work with network and local information:
  - WMS load
  - KML file load
  - GML file load
  - SHP file load
- Geometry drawing and editing tools.
- Local Nodes Remote Administration (BackOffice). Remotely enables management of information from different local nodes.
- Tools for citizen participation.

- I + D developments. In this workgroup there have been developments in these two fields:
    - Environmental monitoring (SOS), responsible for collecting the measurements made by each of the sensors that comprise a network of environmental monitoring, and also for processing and publishing the resulting data.
    - Transformation of OTALEXC GI into WEB 2.0 (linked data and WEB semantic).
- Network SI-OTALEX C (fig. 6)

Figure 5: Cartography harmonization

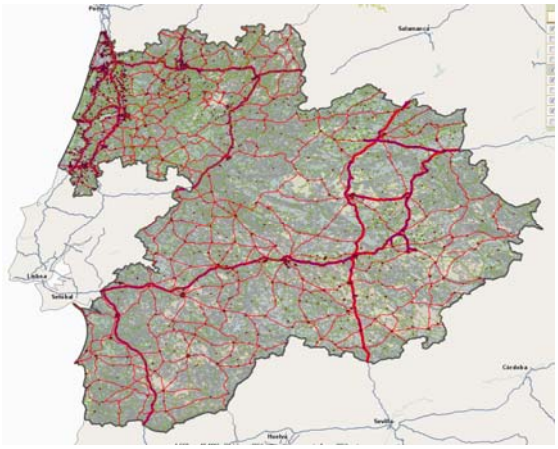
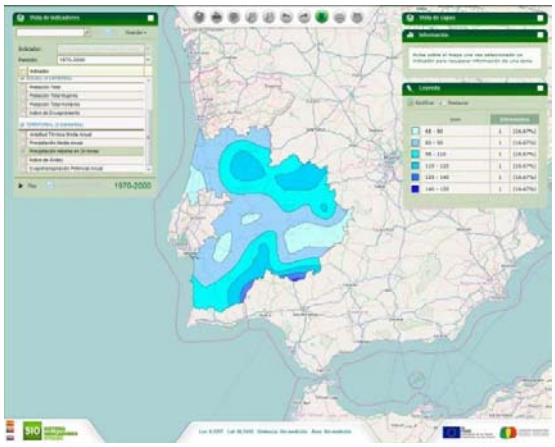


Figure 6: Network SI-OTALEX C



4 References

[1] OECD. OECD core set of indicators for environmental performance reviews. OECD Environment Monographs 1993, No. 83.

[2] IISD. Indicators for Sustainable Development: Theory, Methods, Applications. A Report to the Balaton Group. International Institute for Sustainable Development, Hartmut Bossel (Ed.), 1999.

[3] T. Batista. Spatial Data Infrastructures - key issue for territorial cooperation in Europe: IDE-OTALEX - Alentejo and Extremadura Territorial and Environmental Observatory, Parliament Magazine's, Regional Review Open Days, 2009, 14: 135.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Noise map: professional versus crowdsourced data

Andrea Pődör
University of West Hungary Faculty of Geoinformatics
Pirosalma str.1-3.
Székesfehérvár, Hungary
pa@geo.info.hu

András Révész
Calderdale MBC
Mulcture House,
Mulcture Hall Road
Halifax, HX1 1UN,
UK
andras.revesz@calderdale.gov.uk

Abstract

The goal of the recent study is to evaluate the usability of the data measurement capability of an average smartphone and make a comparative study on available open source mobile applications potentially suitable to noise mapping. In the study a dataset generated by professional equipment was used as a reference. The study confirmed that the mobile applications running on the smartphones tested are not capable for scientific measurement although correlation suggest that calibration may lead to reasonably accurate noise level capture. The study also revealed that different mobile applications produce different outputs. These type of user generated noise measurements cannot substitute professional surveys but can contribute to noise monitoring to testing the effect of the action plans created by the settlements in order to reduce noise pollutions.
Keywords: noise mapping, crowdsourcing.

1 Introduction

Noise pollution is one of the main and growing environmental problems in urban areas. It affects everyday life, well-being and it can even cause severe psychological problems.
According to the Environmental Noise Directive of the European Union 2002/49/EG article 7, agglomerations with a population more than 250 000 should create a noise map.

The Hungarian Government Decree 280/2004. (X.20) in 2004 obliged all settlements to comply with the Directive. The Decree also requires to renew these maps every five years, however due to the lack of financial resources local authorities are unable to fund the renewal of these maps.

An alternative to update the noise maps is crowdsourced data collection. Due to increased availability of location-enabled smart phones with a range of digital sensors including sound recording, this kind of data acquisition is promising.

However it is important to assess the accuracy of surveys carried out by mobile equipment. In this preliminary study we investigated 2 aspects of crowdsourcing noise data collection.

2 Method

We investigated 2 basic phenomenon of noise level survey carried out by commercial smartphones to assess the possibility of crowdsourced noise measurement: (1) we compared the measurements of a professional equipment and a smartphone. (2) We compared the measurements of 2 different mobile applications running on the same model of smartphones.

The standard weighting used in noise measurement is A-Weighting (LA) which applies a frequency dependent weighting on the noise levels according to human perception. A-weighted measurements are expressed as dBA or dB(A).

2.1 Comparing the professional equipment and smartphone

The professional survey was carried out in 2012 in Székesfehérvár by a Brüel & Kjaer sound level meter type 2250 (Class 1) equipment.

We carried out the smartphone survey a year later by Sony Xperia P mobile phones and Sound Meter PRO onboard mobile application. As we had no access to the professional equipment, we repeated the survey when the conditions were the most similar, on the same calendar day of the following year. Both dates fall on a weekday so presumably the traffic conditions were similar and there was no change in the state of other objects responsible for noise emission.

The measurements in the 2 surveys were carried out in the same period of the day averaging 15 seconds of measurement at 51 locations mainly in the inner part of the city. The recorded parameters were (1) lowest time-weighted sound level – LAFmin, (2) higher time-weighted sound level – LAFmax and (3) average sound level – LAeq.

2.2 Comparing different mobile applications

We compared (1) Sound Meter PRO (Smart Tools co.) and (2) Noise meter (JINASYS) mobile applications running on the same models (Sony Xperia P) at the same time and location.

1 minute continuous measurements were averaged starting at 6:00, 11:00, 14:00, 17:00, and 21:00 and repeated 7 times with 10 minutes delays producing 35 parallel measurements. Therefore peak hours and more quiet hours were also

included. The data was recorded at 7-8 m from the axis of the traffic line at the busiest intersection in Gyula, Hungary.

3 Results

There are significant correlations between the professional equipment and the mobile phone measurements in case of LAFmax (p=0.0041) and LAeq (p= 0.0038) but the correlation is not significant in case of LAFmin (p= 0.349963942). The correlation coefficients of LAFmax (r= 0.39) and LAeq (r= 0.40) suggest that there are probable positive correlations between the professional and mobile equipment measurements.
Comparing the results of the professional and mobile phone measurements, the latter produced higher values of dB.

There are also significant correlations between the 2 different mobile application measurements in case of LAmin (p=8.3E-09) and LAeq (p= 0.0097) but not in case of LAmax (p= 0.075). The correlation coefficients of LAmin (r= 0.79) and LAeq (r= 0.43) suggest that there are probable positive correlations between the measurements of the 2 mobile applications. The deviation between the values were approximately 10-15 dB.

4 Conclusions

The mobile application running on the smartphone is not capable for professional survey, however significance of correlations of LAFmax and LAeq suggests that at least change detection is possible which can provide additional data for strategic noise monitoring to test the effect of the action plans created by the settlements in order to reduce noise pollutions.

The higher values of the professional equipment is most likely due to the lack of calibration and the characteristic of the sensors. In a further survey calibration can be tested.

The study also revealed that different mobile applications produce different outputs. As the same sensors were used the difference is probably due to the difference in the calculations by the applications.

The mobile applications retrieve a recorded sample from the microphone. The application can calculate the sound level values from the sample only. Depending on the functions the applications calculates the noise levels from the sound sample, the values may be different. There are many applications available but the second finding suggests that in a crowdsourced survey the same application have to be used by all surveyors.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

SDI strategic planning using the system dynamics technique: A case study in Tanzania

Ali Mansourian, Alex Lubida, Ehsan Abdolmajidi, Petter Pilesjö, Micael Runnström
GIS Center, Department of Physical Geography and Ecosystem Science, Lund University, Sweden
Sölvegatan 12, 22362, Lund, Sweden
{Ali.Mansourian, Alex.Lubida, Ehsan.Abdolmajidi, Micael.Runnström}@nateko.lu.se,
Petter.Pilesjo@gis.lu.se

Abstract

Development of spatial data Infrastructure (SDI) is a long term process, which requires long-term plans. The complexity of SDI, which is a matter of technical, institutional and financial challenges and their interactions, makes the development of such a plan complicated. It is also generally hard to convince policy-makers about the reliability of a plan and the future effect of that to get their supports. The system dynamics technique has been shown to be a proper approach for SDI planning, responding to the above issues. This paper summarizes the application of the system dynamics technique for SDI modelling in Tanzania.

Keywords: Spatial data infrastructure (SDI), the system dynamics technique, strategic planning, Tanzania.

1 Introduction

Spatial data infrastructure (SDI) is typically defined as a set of interacting institutional, technological, human and economic resources that are available for facilitating and coordinating spatial data access, use and sharing. The development of SDI in Tanzania is at the conception stage, with the preparation of a strategic plan as the ongoing step. However, preparing such a plan is a challenging task due to two main reasons. First, SDI has a dynamic complex nature [6,4] due to the variety of interactive and dynamic factors affecting the development of SDI. To prepare a development plan these factors and their interactions and feedbacks on each other should be considered and modelled. Second, the concept of SDI is still new (for inexpert) and evolving. As a result, receiving the support of policy-makers on a plan is generally difficult, where they have no insight about the results of investment on a long-term plan. Such a problem is more highlighted in a developing country, like Tanzania, where has generally limited financial resources.

Mansourian and Abdolmajidi [5] considered SDI as a complex adaptive system and then modelled the development of SDI, using the system dynamics technique, by considering the main affecting factors and their interactions in the relevant community. The model can show the growth (development) of SDI, during the time, based on today's policies. This approach is answering to both the first and the second challenges for SDI planning, mentioned above. With this mind the system dynamic technique was used for SDI strategic planning in Tanzania.

2 The system dynamics technique

System dynamics technique is a method to enhance learning in complex systems and help in designing more effective policies. It involves the usage of feedbacks, flows, states, and time delays to facilitate the integration of the qualitative and quantitative variables operating within the boundary of the system [2,3]. The technique has been widely used in different disciplines including information technology [1]. The system dynamics technique can be used as a virtual world to simulate real situations.

3 SDI modelling in Tanzania

The development of SDI in Tanzania was modeled within three main steps as follow. First, the current status of spatial data activities and plans for the development of national SDI in Tanzania were studied, based on questionnaire survey and interview with members of the Tanzania's SDI steering committee (TSSC). Based on this study, the progress of SDI in Tanzania, with the current situation, was modeled (Figure1)

Then the model was presented to TSSC and refined by receiving feedbacks and comments. Figure 2 shows the progress of SDI within a ten year period. The overall growth of SDI is presented by a Growth index. As Figure 2 shows, within ten years SDI will progress for about 40 percent and then it does not grow any more. The main reasons of such situation were analyzed and documented.

Figure 2: SDI Growth within 10 years, based on a current situation

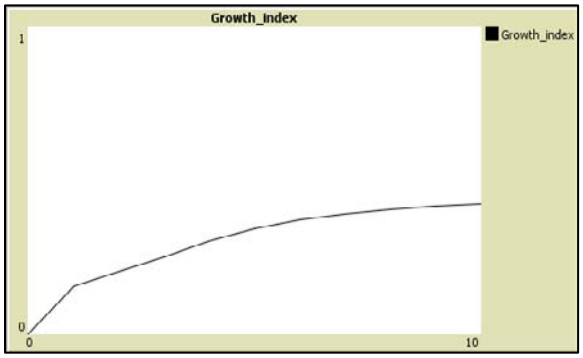
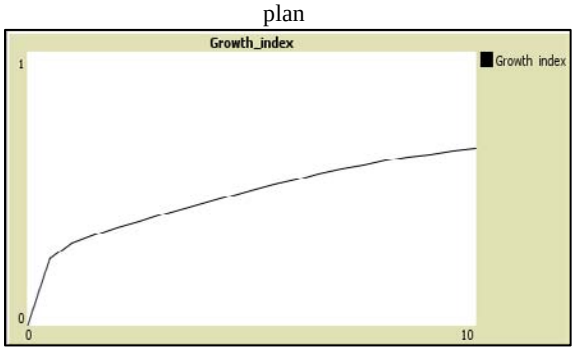


Figure 4: SDI Growth within 10 years, based on the proposed plan



4 Conclusion

The results of the SDI development model in Tanzania were presented to TSSC in a meeting. The approach and the results were interesting for TSSC since they could clearly realize the future effect of their today's decisions on the progress of SDI.

By changing parameters and simulating the development of SDI for different scenarios, it was realized which factors should be considered with high priorities in Tanzanian SDI strategic planning.

References

[1] Dai, X., Xiao, J.H., Xie, K., 2009. Growth of Enterprise Information Technology Application: System Dynamics Model and Empirical Evidence, 27th International Conference of the System Dynamics Society, 26-30 July 2009, Albuquerque, New Mexico, USA.

[2] Dudley, R., Soderquist, C., 1999. A simple example of how system dynamics modeling can clarify and improve discussion and modification of model structure. the 129th Annual Meeting of the American Fisheries Society, 29 August-2 September 1999, Charlotte, North Carolina.

[3] Forrester, J.W., 1980. Information Sources for Modeling the National Economy. Journal of the American Statistical Association, 75(371), 555-574.

[4] Grus, L., J. Crompvoets, and A. K. Bregt, 2010, Spatial Data Infrastructures as Complex Adaptive Systems, International Journal of Geographical Information Science 24(3), 439–463.

[5] Mansourian, A., Abdolmajidi, E. (2011). Investigating the system dynamics technique for the modeling and simulation of the development of spatial data infrastructures, International Journal of Geographical Information Science 25, no. 12 (2011), 2001–2023.

[6] Rajabifard, A., Feeney, M-E.F., Williamson, I.P., 2002. Future directions for SDI development. International Journal of Applied Earth Observation and Geoinformation. 4(1), 11–22.

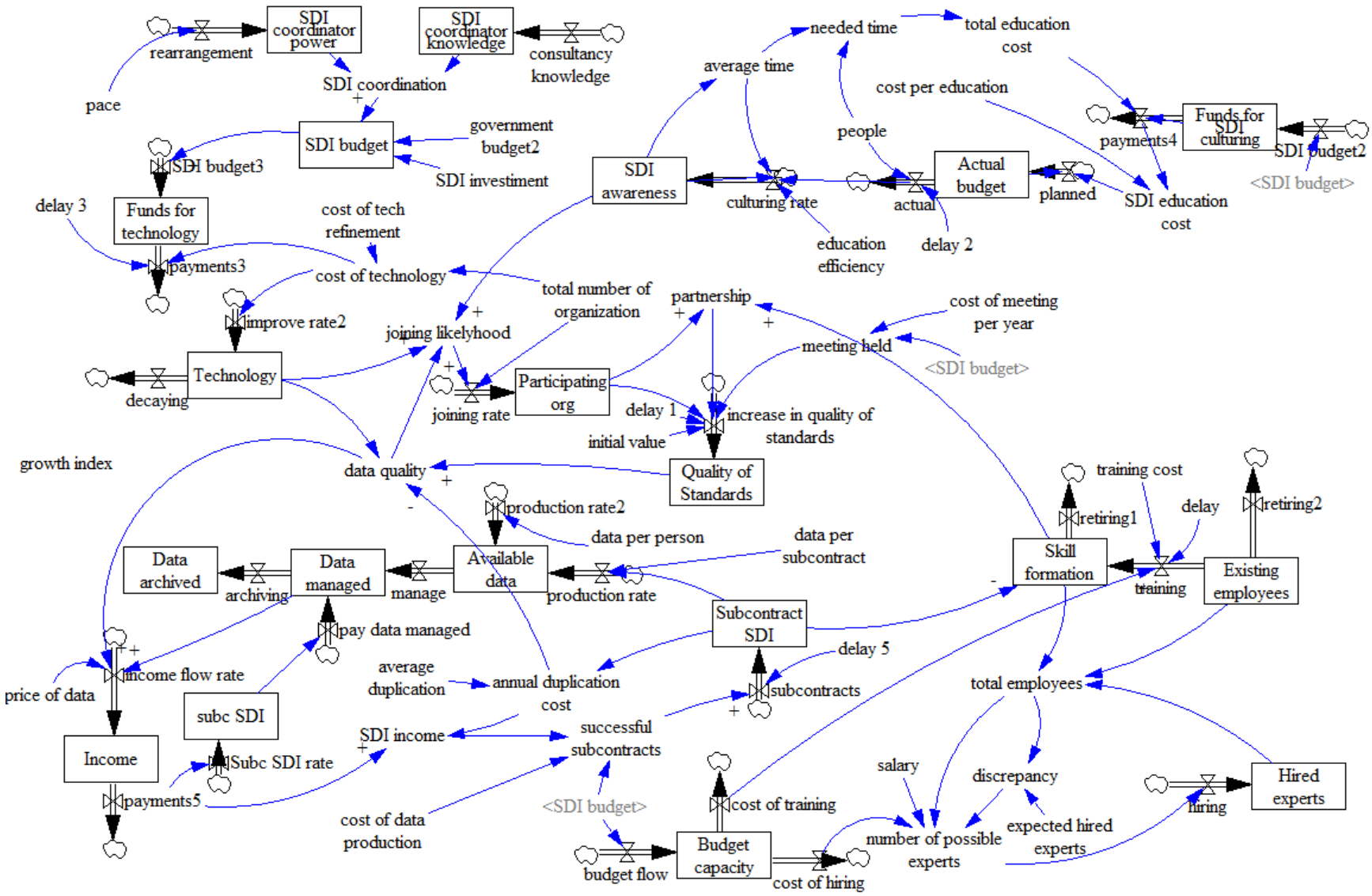


Figure 1: A system dynamic model developed based on the current status of spatial data activities and SDI plans in Tanzania.

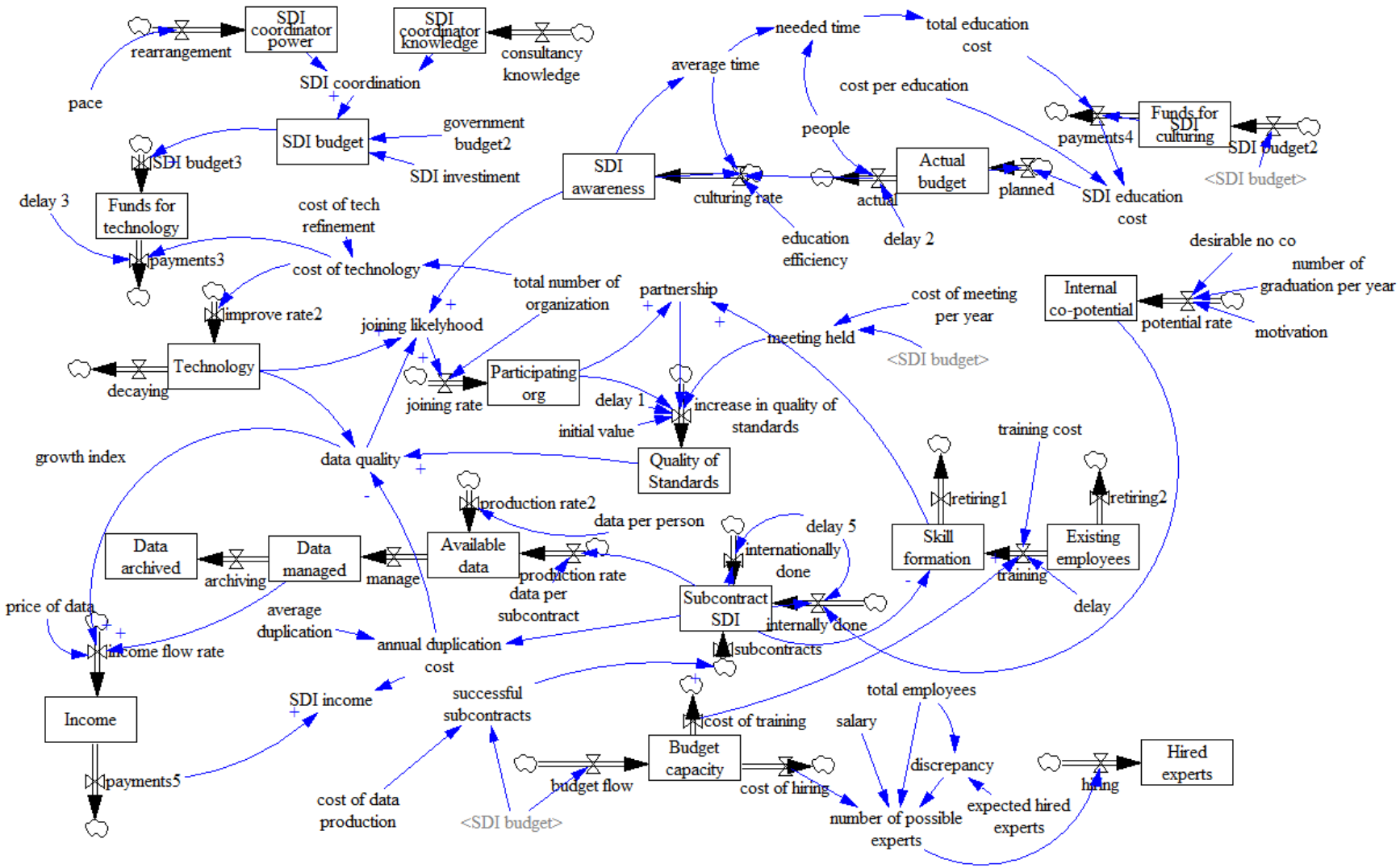


Figure 3: A system dynamic model presenting the proposed plan for developing Tanzania SDI.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

A conceptual representation for modelling the synchronization process of complex road networks

M.Dolores Arteaga
University of New Brunswick
15 Dineen Drive
Fredericton, Canada
md.arteaga@unb.ca

Monica Wachowicz
University of New Brunswick
15 Dineen Drive
Fredericton, Canada
monicaw@unb.ca

Keywords: large graph network, synchronization, complex road network, transportation.

The conceptual representation of complex road networks goes beyond the established dualism of spatio-temporal database structures in GIS (e.g. raster versus vector, continuous versus discrete view of time) and moves further towards a more topological structure where a synchronization process affects the probability of linking the nodes of a complex network [1]. Predicting whether dense traffic will congest involves issues with respect to this synchronization process. In particular, Manson [2] provides an interesting effort at describing the evolution of complexity research and its application in the geographical domain.

Synchronization is a process wherein complex road networks adjust a given property of their motion due a suitable coupling of their structures (i.e. nodes and links) to an external forcing driven by the behavior of the same networks [3]. It challenges the development of new methods in complex road networks to support real-time transportation analytics for discovering patterns in traffic networks and revealing new insights to decision making.

Most of the existing representations developed in the transportation domain are graph-based networks where nodes and links are used to represent the topological and geometric properties of the intersections and the roads of a network [4, 5]. In the primal graph-based networks, intersections are represented by nodes and roads are represented by links [6]. This representation has been very useful to understand the connectivity between roads through representing the distance of their consecutive links. Conversely, in the dual graph-based networks, the roads are represented by nodes and intersections are represented by links [7]. This representation has played an important role in the computation of the accessibility and integration within road networks. In both representations, statistical properties, such as centrality measures, clustering or cellular structure, have been used to analyze patterns in traffic networks [8, 9, 10]. However, these properties only provide a snapshot of the behavior of a road network rather than its dynamic synchronization behavior [11].

In this poster presentation, we demonstrate how the coupling of a topological structure to its behavior is achieved using a large graph-based representation where nodes are all the intersections found within a complex road network as well as links represent the synchronization between these nodes (i.e. intersections). The synchronization properties are coupled to the links for modeling the motion of vehicles between the intersections. The research challenge is twofold:

- The organization of intersections is not completely represented by a unique hierarchy, but by a set of several hierarchical levels that appear at different topological scales [12]. For example, intersections might be the location of highway exists or traffic lights.
- Highly densely connected sets of nodes synchronizes more easily that those with sparse connections. Therefore, for complex road networks with poor connectivity, starting from random initial measures, those highly interconnected nodes forming local clusters will synchronize first. This can be illustrated by the principle which 20 percent of roads accommodate 80 percent of traffic flow (20/80) [9].

Our approach differs from previous work in proposing synchronization for representing the dynamic behavior of a complex road network because it is the whole process which will reveal the topological structure at different scales.

References

[1] M. Wachowicz and J. B. Owens. Space-time representations of complex networks: What is next. Geofocus 9(1), pp. 1-8. 2009.

[2] S. M. Manson. Simplifying complexity: A review of complexity theory. Geoforum 32(3), pp. 405-414. 2001.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. Hwang. Complex networks: Structure and dynamics. Physics Reports 424(4), pp. 175-308. 2006.

[4] M. Rosvall, A. Trusina, P. Minnhagen and K. Sneppen. Networks and cities: An information perspective. Phys. Rev. Lett. 94(2), pp. 028701. 2005.

[5] S. Gao, Y. Wang, Y. Gao and Y. Liu. Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. Environment and Planning B: Planning and Design 40(1), pp. 135-153. 2013.

[6] S. Porta, P. Crucitti and V. Latora. The network analysis of urban streets: A primal approach. ArXiv Preprint physics/0506009 2005.

[7] S. Porta, P. Crucitti and V. Latora. The network analysis of urban streets: A dual approach. Physica A: Statistical

Mechanics and its Applications 369(2), pp. 853-866. 2006

[8] J. Lin and Y. Ban. Complex network topology of transportation systems. Transport Reviews 33(6), pp. 658-685. 2013.

[9] B. Jiang. Street hierarchies: A minority of streets account for a majority of traffic flow. Int. J. Geogr. Inf. Sci. 23(8), pp. 1033-1048. 2009.

[10] I. Ntoutsi, N. Mitsou and G. Marketos. Traffic mining in a road-network: How does the traffic flow? International Journal of Business Intelligence and Data Mining 3(1), pp. 82-98. 2008.

[11] A. Arenas, A. Díaz-Guilera and C. J. Pérez-Vicente. Synchronization processes in complex networks. Physica D 224(1–2), pp. 27-34. 2006.

[12] A. Arenas, A. Díaz-Guilera and C. J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. Phys. Rev. Lett. 96(11), pp. 114102. 2006.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Evaluation of subjective preferences regarding indoor maps: comparison of schematic maps and floor plans

Luciene Stamato Delazari
Geodetic Science Program
Federal University of Paraná
Curitiba, Brazil
luciene@ufpr.br

Suchith Anand, Jeremy Morley
Nottingham Geospatial Institute
University of Nottingham,
Nottingham, UK
{suchith.anand,
jeremy.morley}@nottingham.ac.uk

Abstract

In this study, we investigate subjective preferences regarding floor plans and schematic maps in the use of a map in an indoor environment. To achieve this, we performed a qualitative experiment with a random user sample; the survey was carried out remotely. The survey was conducted in Portuguese and English, and users were asked to answer questions, using two different maps:a floor plan and a schematic map. In the sequence, users were asked questions about their preferences regarding map use in an indoor environment. Users also answered questions about the positive and negative aspects of using a schematic map in an indoor environment. The initial results do not indicate a preference for one kind of map, but show that users found the symbology adopted in the schematic map easier to understand.
Keywords: Indoor maps; schematic maps; subjective preference

1 Introduction

The growth in the size and complexity of public buildings such as universities, airports, and shopping malls has made efficient indoor navigation necessary. Examples of indoor navigation tools are "You Are Here" (YAH) maps. The main objective of YAH maps is to aid navigation, but there are issues concerning their use, such as misalignment [1], object rotation, and self-location [2].

Schematic maps (SMs) are helpful in spatial problem-solving tasks such as way-finding in outdoor environments, or for representing underground railways, surface railways, and tram and bus routes. There has been significant research on methods for obtaining a schematic representation from a topological structure [3, 4, 5, 6]. Research on indoor maps is more recent, and has focused on positioning techniques rather than the representation of such spaces [7].

There is no established knowledge regarding which type of map is best for an indoor environment; therefore this paper presents a study of subjective preferences, comparing an FP and an SM of two different buildings.

2 Methodology

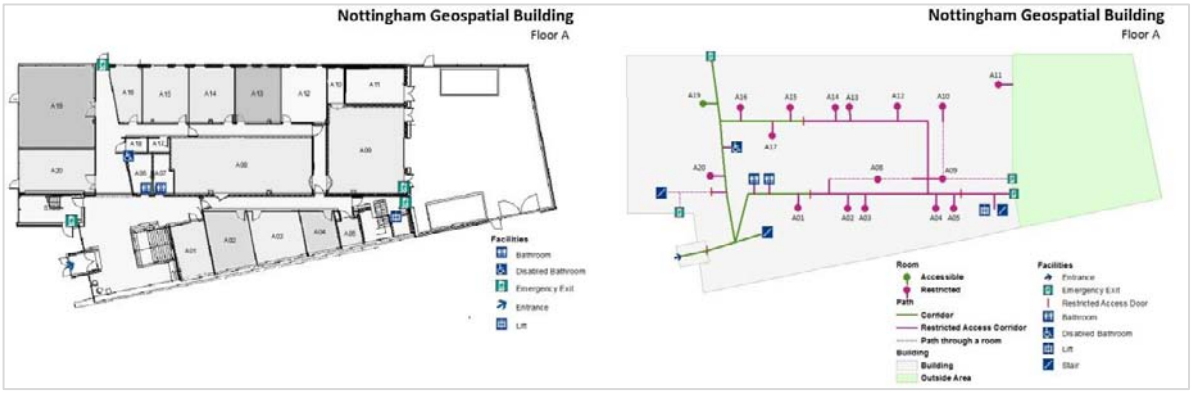
The first step in this research was to understand how to design an SM for an indoor environment. After the map was designed, we developed a survey to compare the preferences of users regarding FPs and SMs.

2.1 Schematic Map Design

We used a floor of the Nottingham Geospatial Building and a floor of the Portland Building, both at the University of Nottingham (UK). For each floor, there were an FP and the proposed SM.

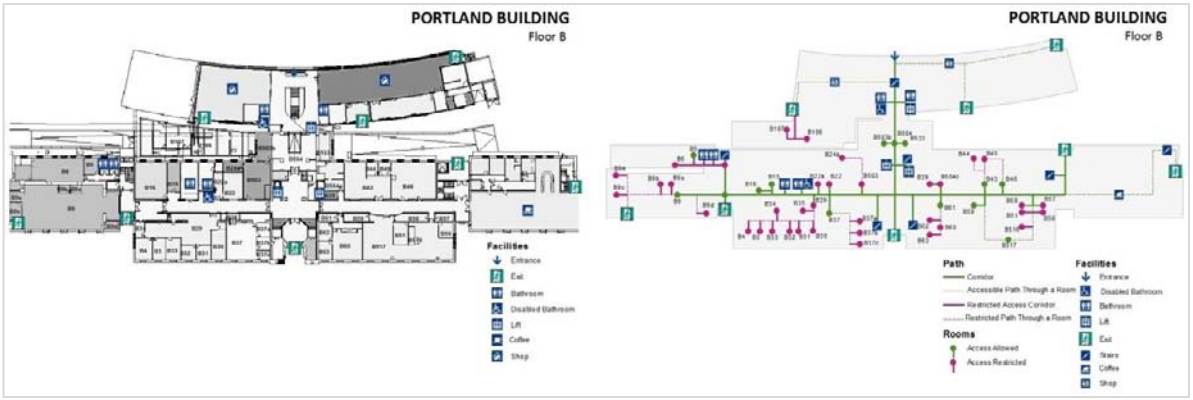
The base map is composed by the building external walls and possible subclasses; the thematic data are as follows**:** **c**orridors, rooms, interest points. The three classes of objects defined in the map design have specific rules. Thus, rooms are generalized to points, which are linked to paths by lines at the door positions. All interest points are represented by pictorial symbols representing the original objects. Paths are lines between entrance and exit points, connecting the rooms and interest points. Lines connecting adjacent rooms are represented by solid lines, and lines connecting rooms inside other rooms, or restricted access areas, are represented by dashed lines. The FPs and respective SMs are presented in Figures 1 and 2

Figure 1 - Nottingham Geospatial Building: (a) floor plan; (b) schematic map



(a) (b)

Figure 2 - Portland Building: (a) floor plan; (b) schematic map



(a) (b)

2.2 Experiment

The experiment was developed as an online survey in Portuguese and in English (www.cartografia.ufpr.br/indoor_test/survey_indoor.php). The general design of the survey is presented in Figure 3. After the user chose the language, he/she was randomly assigned to Group A or B and, in both cases, had to provide some personal information. If the user was assigned to Group A, the order of map presentation was the FP and then the SM. This order was reverse for Group B. In both cases, the users had to answer two questions about each map. The user was instructed to observe the map carefully before starting the survey. In the sequence, we presented the same map, but some symbols were changed or removed. The questions were as follows:

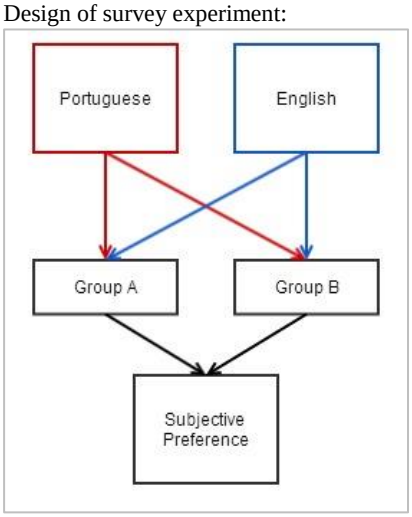
- NGI Building:

a) Please identify the position of the Lift.
b) You are at the position marked in the map and there is a fire drill. Which is the BEST path from your position to an Emergency Exit?

- Portland Building:

a) You need to inform a person in a wheel chair where there is a Disabled Bathroom. Identify on the map the locations of these bathrooms.
b) You are at the position indicated and there is a fire drill. You have to find one of the emergency exits. Which is the VALID Emergency Exit represented in the map?

Figure 3 - Design of survey experiment:



The user answered a set of questions about his/her preferences regarding the use of maps in an indoor environment. The questions were as follow:

P1 Which map did you find easier to use in order to learn about an indoor environment?
P2 Which map did you think was easier to understand?
P3 When comparing Schematic Maps and Floor Plans, which type of map would you prefer to use in an indoor environment?
P4 Please say how difficult or easy you found it to understand the SM map (options: very easy, easy, difficult, and very difficult).
P5 In your opinion, what is a positive point regarding using a Schematic Map to represent an indoor environment (options: easy to understand, symbology, simplicity, and none).
P6 In your opinion, what is a negative point regarding using a Schematic Map to represent an indoor environment compared with a Floor Plan? (options: difficult to understand, symbology, complexity, and none).
P7 Do you think it is important to represent on the map where you have to cross one room to get to another?
P8 What symbology do you prefer to represent Rooms on a Schematic Map? In this case, two SMs were presented; one is the same as in Figure 1, and the other had squares instead circles for representing rooms.

User tests were collected using *HTML* forms, filled in via the web, and stored in a server, using a short *PHP* script. Data were received in text format and inserted into an Excel table. Tables related to the tests were built separately for each language, including a table giving user characteristics. Each user received a random number identifier to remove any possibility of identification.

3 Results and future work

We received 140 answers, divided into 93 Portuguese and 47 English users. In these two groups, most users were female. The Portuguese users were mostly educated to undergraduate level, and English users mostly had master's degrees. In both groups, the majority stated that they often use maps and around 50% of users reported that they sometimes look for indoor maps.

Table 1 presents the answers regarding subjective preferences (questions P1 to P3). Questions P4, P5, and P6 were related to the SM only. In these questions, users were asked to rank how easy or difficult they found it to understand the map, and to state positive and negative points regarding using an SM to represent an indoor environment compared with an FP. Finally, there were two questions about the map symbology (P7 and P8). The results for these questions are presented in Table 2.

Based on these initial results it is not possible to state that one type of map is preferred by users, but this is an important step in helping understand the usage of schematic maps in indoor environments. With regard to the questions about the SM only, both groups found it easy to understand and pointed out the simplicity as a positive point. The percentage of users that said that they preferred the FP is too small for us to report that FP are better for representing an indoor environment. Furthermore, there are preference differences when analyzing the groups separately. Further investigations will consider both groups in order to confirm this preference.

Future work will be done by testing users in real situations to determine whether Schematic Maps can be useful tools for helping with way-finding tasks.

Table 1: Results for questions about use preferences (in %)

	Portuguese		English		Total	
	FP	SM	FP	SM	FP	SM
P1	48.4	51.6	67.4	32.6	54.7	45.3
P2	52.7	47.3	67.4	32.6	57.6	42.4
P3	47.3	52.7	67.4	32.6	54.0	46.0

Table 2: Results for questions related to schematic map (in %)

	Port.	**English**	**Total**
P4			
Very easy	10.8	10.9	10.8
Easy	64.5	54.3	59.4
Difficult	24.7	26.1	25.4
Very difficult	-	8.7	4.3
P5			
Easy	6.5	6.5	6.5
Simplicity	55.9	45.7	50.8
Symbology	21.5	30.4	26.0
None	16.1	17.4	16.8
P6			
Complexity	35.5	41.3	38.4
Difficulty	8.6	30.4	19.5
Symbology	20.4	15.2	17.8
None	35.5	13.0	24.3
P7			
No	29.0	23.9	26.5
Yes	71.0	76.1	73.5
P8			
A	49.5	47.8	48.6
B	50.5	52.2	51.4

4 Acknowledgement

This research is funded by CAPES. We also wish to thank all the volunteers that participated in the survey.

5 References

[1] D. R. Montello, "You Are Where? The Function and Frustration of You-Are-Here (YAH) Maps," *Spat. Cogn. Comput.*, vol. 10, no. 2–3, pp. 94–104, Jun. 2010.

[2] A. K. Lobben, "Tasks , Strategies , and Cognitive Processes Associated With Navigational Map Reading : A Review Perspective," *Prof. Geogr.*, vol. 56, no. August 2013, pp. 270–281, 2004.

[3] S. Avelar and L. Hurni, "On the Design of Schematic Transport Maps," *Cartogr. Int. J. Geogr. Inf. Geovisualization*, vol. 41, no. 3, pp. 217–228, Sep. 2006.

[4] J. M. Ware, G. E. Taylor, S. Anand, and N. Thomas, "Automated Production of Schematic Maps for Mobile Applications," *Trans. GIS*, vol. 10, no. 1, pp. 25–42, Jan. 2006.

[5] D. Weihua, G. Qingsheng, and L. Jiping, "Schematic road network map progressive generalization based on multiple constraints," *Geo-spatial Inf. Sci.*, vol. 11, no. 3, pp. 215–220, Jan. 2008.

[6] D. Weihua, L. Jiping, and G. Qingsheng, "Visualizing schematic maps through generalization based on adaptative regular square grid model," in *The International Achives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2008, pp. 379–384.

[7] A. Puikkonen, A. Sarjanoja, M. Haveri, J. Huhtala, and J. Häkkilä, "Towards Designing Better Maps for Indoor Navigation – Experiences from a Case Study," in *MUM´09*, 2009, pp. 1–4.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Using Crop Phenology to Assess Changes in Cultivated Land after the Anfal Genocide in Iraqi Kurdistan

Lina Eklund
Lund University, Sweden

Andreas Persson
Lund University, Sweden

Jing Tang
Lund University, Sweden

Mitch Selander
Lund University, Sweden

Martin Borg
Lund University, Sweden

Abstract

The Anfal genocide campaign, carried out by the Iraqi government against the Kurdish population in 1988, has been reported to have severe consequences for agriculture and food security by causing large scale land abandonment.

This study uses Landsat satellite data to detect agricultural changes that can be attributed to the Anfal genocide. Cultivated land were distinguished from other land cover types by focusing on crop phenology.

Initial results show a strong decrease in cultivated land in the years after the genocide, especially in the areas that were targeted by the genocide campaign.

Keywords: Agriculture, Crop phenology, Genocide, Iraqi Kurdistan, NDVI

1 Introduction

During the early 1980s, the northern Kurdish governorates accounted for between 25 and 30% of the total Iraqi food production [1]. The Anfal genocide campaign, carried out as a series of attacks on the Kurdish areas in 1988, included bombings of villages, destruction of agricultural land, and displacement from villages [2, 4]. This caused a drastic change from a mainly rural to an urban population.

Effects of conflicts on land have been studied satellite images and geographic analyses (cf. Kuemmerle *et. al* [5] and Gibson *et. al* [6]). Mubareka and Ehrlich [3] focused on land use changes in Anfal affected areas and found that agricultural land in the Sulaymaniah governorate had decreased, but not in the Duhok governorate. This, however, contradicts other reports of the Anfal effects on agriculture in Kurdistan [2-4], which motivates further analysis

This paper aims at quantifying the loss of agricultural land in connection to the Anfal genocide in the Duhok governorate, by using multi-temporal high resolution satellite imagery.

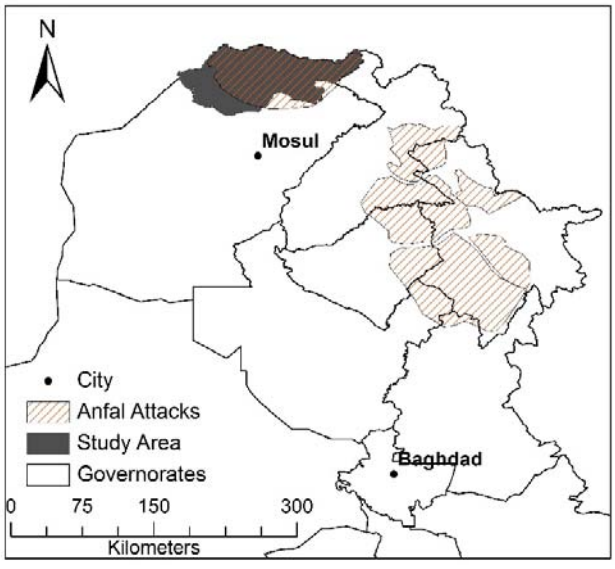
2 Data and Methods

Study Area

The Duhok governorate belongs to the Kurdistan Region, which is a semi-autonomous region situated in northern Iraq (figure 1).

In the plains, the main agriculture is cereal production, such as barley and wheat, which is harvested in June. In the mountain areas, orchards with fruit, nuts and almonds are common.

Figure 1: The study area location in northern Iraq and the Anfal genocide attack.



Data

This assessment uses data from two periods: pre-Anfal and post-Anfal (table 1). Seventeen surface reflectance images recorded by the Landsat 4 and 5 Thematic Mapper (TM) satellites were used. In addition to satellite data, digitized spatial data of the Anfal affected areas, and basic geographic data of administrative borders and cities were used for analysis and visualization.

Table 1: Julian dates for the land surface reflectance images.

PRE-ANFAL					
1984		1986		1987	
Spr.	Sum.	Spr.	Sum.	Spr.	Sum.
158		115		22	198
		163		150	246
POST-ANFAL					
1989		1990		1991	
Spr.	Sum.	Spr.	Sum.	Spr.	Sum.
35	195	86	190	129	
	211	134	238		
	243		254		

Methods

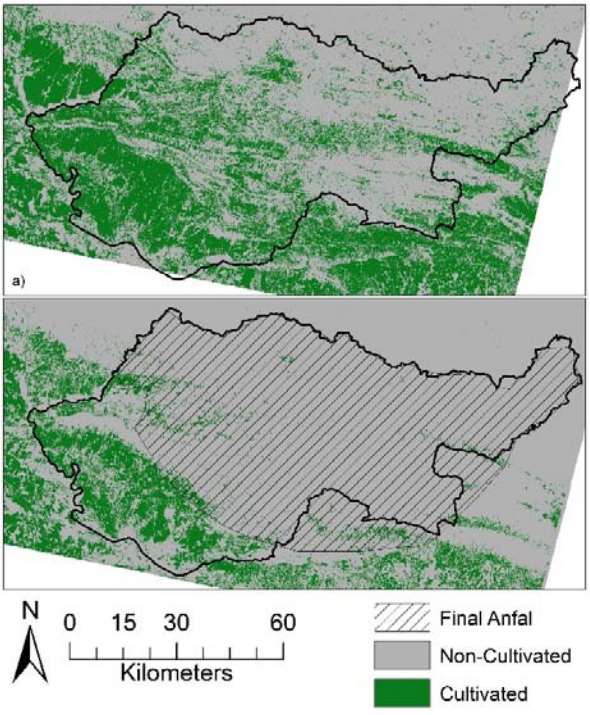
Crop Phenology is based on the different inter-annual variability in greenness of different land covers [6]. Agricultural cropland exhibit a large difference in NDVI between pre- and post-harvest months.

Normalized Difference Vegetation Index (NDVI) were calculated for all images and Maximum value composites (MVC) for pre-harvest and post-harvest periods were calculated for both periods. From the MVC images, ranges were calculated. Pixels with ranges larger than 0.3 indicated harvest and were classified as cultivated land.

3 Results

Much of the cultivated land had decreased in the post-Anfal period, especially in the areas that were targeted by the final Anfal attack (figure 3). Changes are, however, seen in the whole area, even in Turkey in the North West.

Figure 3: Areas in the Duhok governorate classified as cultivated land in the a) pre-Anfal period and b) post-Anfal period.



A decrease of 1438 km^2 in cultivated land were recorded in the whole governorate, representing 62% of the pre-Anfal cultivated area (table 2). In the Anfal area, cultivated land had decreased from 1283 km^2 to 150 km^2, representing a decrease of 88% of the original cultivated area.

Table 2: The change in cultivated land area within the Duhok governorate and the area targeted by Anfal

DUHOK GOVERNORATE				
	Pre Anfal (km^2)	Pre Anfal (%)	Post Anfal (km^2)	Post Anfal (%)
Cultiv.	2311	35	873	13
Other	4251	65	5706	87
Sum	6562	100	6579	100
ANFAL AREA				
	Pre Anfal (km^2)	Pre Anfal (%)	Post Anfal (km^2)	Post Anfal (%)
Cultiv.	1283	26	150	3
Other	3673	74	4815	97
Sum	4956	100	4964	100

4 Discussion

The results indicate a strong decline in cultivated land area after the Anfal attack in 1988, when many people were killed and forcibly moved, and about 4000 villages were destroyed [3]. Before Anfal, the Duhok governorate was a mainly rural area where livelihoods were largely based on agriculture [2]. The displacement of people during and after the Anfal attacks, together with the destruction of villages and lands, led to a decrease of more than 60% in cultivated areas. The change was especially severe in the mountain areas that were exposed to the attacks.

Several reports state that the Anfal genocide had negative consequences for agriculture in the Kurdistan Region, but until now it has not been proven quantitatively at province level. Using crop phenology of multi-temporal satellite images is a good alternative for areas and periods where agriculture is small scale and data is scarce.

5 References

[1] World Food Programme Iraq - North Coordination Office. *"Oil For Food" - Food Basket Adequacy Assessment Survey* (draft), 2001.

[2] C. Hardi. *Gendered experiences of genocide: Anfal survivors in Kurdistan-Iraq.* Ashgate Publishing Company, Great Britain 2011.

[3] S. Mubareka and D. Ehrlich. Identifying and modelling environmental indicators for assessing population vulnerability to conflict using ground and satellite data. *Ecological Indicators*, 10(2):493-503, 2010.

[4] Human Rights Watch. *Genocide in Iraq - The Anfal Campaign Against the Kurds*, New York 1993.

[5] T. Kuemmerle, D. Müller, P. Griffiths and M. Rusu. Land use change in Southern Romania after the collapse of socialism. *Regional Environmental Change*, 9(1): 1-12, 2009.

[6] G.R. Gibson, J.B. Campbell, and R.H. Wynne. Three decades of war and food insecurity in Iraq. *Photogrammetric Engineering and Remote Sensing* 78(8): 885-895, 2012.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Fuzzy viewshed, probable viewshed, and their use in the analysis of prehistoric monuments placement in Western Slovakia

Alexandra Rášová
Slovak University of Technology in Bratislava
Radlinskeho 11
Bratislava, Slovakia
alexandra.rasova@stuba.sk

Abstract

Viewshed analysis is used in many fields (e.g. landscape architecture, military, or archaeology) to determinate locations visible from one or more observation points in order to examine suitability of the placement of structures or their visual impact on the environment. The output of this analysis in the GIS environment is usually a binary raster with cells coded as "visible" or "invisible". There are several factors that affect the visibility calculation; in this work we address two of them: (i) the effect of the uncertainty of the DEM, (ii) the non-binary nature of human visual perception. We created two toolboxes in ArcGIS Model Builder: "Probable Viewshed" to consider the vertical error of a DEM, and "Fuzzy Viewshed" to assess the changing visibility of an object due to its size and distance from the observer. Both probable and fuzzy viewshed represents visibility as a value from the interval from 0 to 1. This value represents the probability of a cell being visible considering the vertical error of the DEM for the probable viewshed; for the fuzzy viewshed it can be interpreted as the level of clarity of visibility. We used these tools in an archaeological analysis of seven circular ditched enclosures ("roundels") in Western Slovakia. The results confirmed mutual visibility of two quadruples of roundels, so visibility could be one of the determining factors of their placement.
Keywords: GIS, visibility analysis, fuzzy viewshed, probable viewshed, circular ditched enclosure, roundel

1 Introduction

Visibility analysis seems to be an easy task using built-in tools in a GIS. Only few clicks are needed to get the map of visible and invisible areas, but some issues should be considered to acquire more realistic results, e.g. the ambiguous nature of visibility, which cannot be expressed by binary "visible"/"invisible", and the effect of the inaccuracy of a digital elevation model (DEM).

These problems can be addressed using non-binary viewsheds. We created two toolboxes in ArcGIS ModelBuilder: "Probable viewshed", which calculates the probability of visibility of a cell considering the vertical accuracy of a DEM, and "Fuzzy viewshed", which uses a membership function to assign the value of visibility according to the distance and size of an observed object.

Created toolboxes were used in the archaeological analysis of the placement of prehistoric monuments, circular ditched enclosures (roundels). The analysis of their mutual visibility could partially explain their unknown function.

2 Experimental

2.1 Location and data

The analyzed Neolithic circular ditched enclosures ("roundels") are located in Western Slovakia, information about them was provided by Slovak Academy of Sciences, Institute of Archaeology. From overall seven objects, three have been confirmed by geophysical measurements (Prašník – „P", Šterusy – „S", Borovce – „B") and four have been identified from aerial photographs only ("assumed" roundels: Borovce 2 – "B2", Vrbové – "V", Trebatice – "T", Kočín – "K").

The DEM used in this study was the digital terrain model (DTM) with 10 m resolution. Absolute accuracy of its vertical component (RMSE = 0.84 m, standard deviation = 0.64 m) was specified in the quality assessment [5].

2.2 Fuzzy viewshed

"Fuzzy viewshed" toolbox that we created is using the membership function:

$$\mu_f = \frac{1}{1 + 2\left(\frac{d - b_1}{b_2}\right)^2} \quad \text{for } d > b_1; \qquad (2)$$

b_1 is the distance of clear visibility, d Euclidean distance, and b_2 critical distance for an object to be recognized by human eye:

$$b_2 = \frac{h}{2\tan(\beta/2)} \quad , \qquad (3)$$

where h is the size of an object (height or width), β is the recognition acuity of human eye. [3, 6]

We computed the fuzzy viewshed to determine the area, where a standing person could be visible, assuming the observer height 1.5 m (the height of eye line) and the size of the target 1.64 m (average height of a Neolithic man [1]).

2.3 Probable viewshed

The computation of probable viewshed is described in [2] using Monte Carlo simulation; it represents the possibility of a cell being visible considering the DEM inaccuracy, which is expressed as a value from the interval [0, 1]. We created the "Probable Viewshed" toolbox using this approach combined

with an option of considering the spatial autocorrelation of the DEM error using low-mean filter, as suggested by [4, 7].

A probable viewshed was calculated for each roundel from 100 random realizations. We used the uniform distribution with low-mean filter. Given the size of roundels and their mutual distance, it wasn't necessary to consider the fuzzy character of visibility in determining the mutual visibility of the objects. To inquire about visibility patterns, multiple probable viewsheds were calculated for two sets of four mutually visible roundels.

3 Results and discussion

The fuzzy viewshed analysis (an example - Figure 1) showed that there is a possibility to recognize a person standing in front of a neighbouring roundel (assuming the roundel itself was visible).

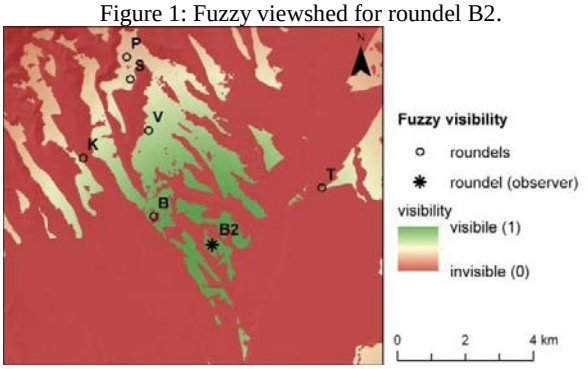
The probable viewshed was used to determine the mutual visibility (Table 1). Two roundels (T, K) are not mutually visible with the others (very low values). In the set of 5 roundels, there are two mutually visible quadruples (P-B-B2-V; S-B-B2-V). P and S, which are only about 660 m distant from each other, are not mutually visible. These two roundels have similar visibility patterns in relation to other 3 roundels (B, B2, V), as can be seen from multiple probable viewsheds (Figure 2, Figure 3). Values of cumulative probability represent sums of single probable viewsheds: a value close to 4 means that this location was probably visible from each observing point. It is thus possible, that one roundel replaced the other because of better position.

Table 1: Probable visibility of the monuments

roundel	B*	P*	S*	B2	V	T	K
B*		1.00	1.00	0.88	0.97	0.00	0.00
P*	1.00		0.20	0.99	0.87	0.99	0.00
S*	1.00	0.40		0.98	0.70	0.00	0.00
B2	0.92	1.00	1.00		1.00	0.25	0.00
V	1.00	0.88	0.70	0.98		0.65	0.00
T	0.00	0.30	0.05	0.25	0.50		0.00
K	0.00	0.00	0.00	0.00	0.00	0.00	

* confirmed roundels

Figure 1: Fuzzy viewshed for roundel B2.



4 Conclusions

Fuzzy viewshed and probable viewshed are tools providing additional information compared to the binary viewshed analysis. Fuzzy viewshed expresses the change of the level of visibility of analyzed object due to its size and distance from an observing point. Probable viewshed provides estimation of the effect of a DEM on calculated visibility; it can be used to confirm that the visibility is not caused by the error of the DEM. We created "Probable Viewshed" and "Fuzzy Viewshed" toolboxes in ArcGIS ModelBuilder; both are published on ArcGIS Resources.

We used these tools to analyze the visibility of prehistoric monuments (roundels). From 7 objects, there are two sets of four mutually visible roundels. The placement of these quadruples of roundels enables to recognize (i) a person standing in the surroundings of at least one neighboring roundel, (ii) all other structures. This suggests possible defense or cult function, because this placement is convenient for signal exchange: to send a warning or participate in a ritual. However, more research is needed to confirm these hypotheses, particularly geophysical measurements to confirm their age and origin and consideration of other factors that affect visibility.

Figure 2: Multiple probable viewshed of roundels P-B-B2-V.

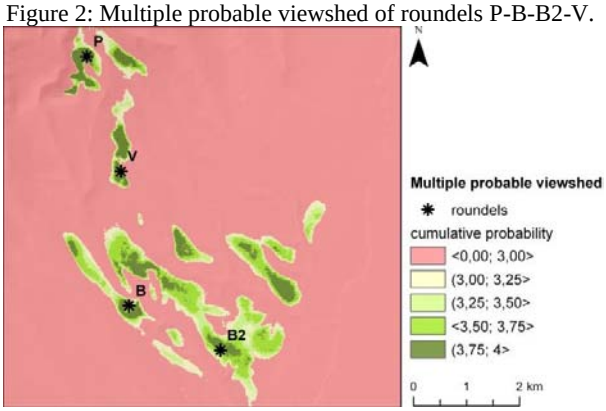
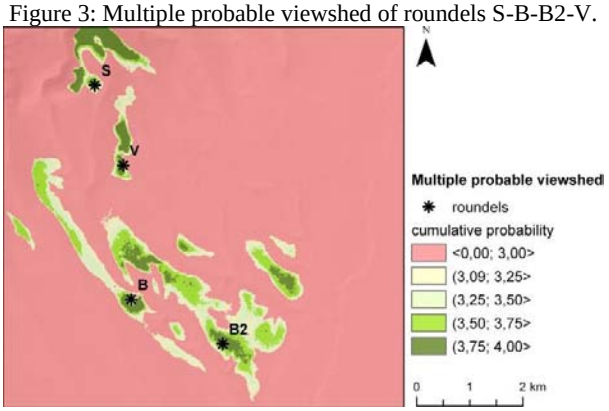


Figure 3: Multiple probable viewshed of roundels S-B-B2-V.



References

[1] D. Alexis, A. Sarris, T. Astaras and K. Albanakis. Integrated GIS, remote sensing and geomorphologic approaches for the reconstruction of the landscape habitation of Thessaly during the Neolithic period. In *Journal of Archaeological Science*, 38: 89-100, 2011.

[2] P. Fisher. First experiments in viewshed uncertainty: simulating fuzzy viewsheds. In *Photogrammetric engineering and remote sensing,* 58: 345-352, 1992.

[3] P. Fisher. Probable and fuzzy models of viewshed operation. In M.F. Worboys, editor. *Innovations in GIS 1: selected papers from the First National Conference on GIS Research UK,* pages 161-175. Taylor, London, 1994.

[4] P. Fisher. Improved Modeling of Elevation Error with Geostatistics. In GeoInformatica, 2(3): 215-233, 1998.

[5] E. Mičietová and M. Iring. Hodnotenie kvality digitálnych výškových modelov. In *Geodetický a kartografický obzor,* 57(99):45-60, 2011.

[6] D.E. Ogburn. Assessing the level of visibility of cultural objects in past landscapes. In *Journal of Archaeological Sciences,* 33:405-413, 2006.

[7] S. Wechsler. Digital elevation model (DEM) uncertainty: evaluation and effect on topographic parameters. In *Proceedings of the ESRI User Conference*, San Diego, 1999.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Usability Patterns for Geoportals

Christin Henzen
Geoinformation Systems, TU Dresden
Helmholtzstr. 10
Dresden, Germany
christin.henzen@tu-dresden.de

Lars Bernard
Geoinformation Systems, TU Dresden
Helmholtzstr. 10
Dresden, Germany
lars.bernard@tu-dresden.de

Abstract

Current geoportals and metadata catalogues, as user interfaces for discovery and exploration for geodata do still suffer from lacking usability, regardless whether experts or non-expert users are considered. Design patterns are well established in software development to tackle frequently occurring problems. Usability patterns are a specialization of such design patterns to specifically address user interface issues and related software solutions. However, existing usability patterns are not sufficient to cope with GI-usability issues as for instance related to discovery of geodata. This poster submission introduces an adapted GI-usability pattern concept.

Keywords: Usability, Pattern, Discovery, Geoportal

1 Motivation

Design patterns describe general reusable solutions to solve frequently occurring problems [1]. In software engineering, this concept has been established and proven for years and extended for several aspects.
Current geoportal implementations still show various and frequent usage issues, e.g. during search and visualization processes. Typical usability problems are linked to the representation of search results (e.g. unclear labelling, irregular result categorization), to the navigation in the result sets (e.g. missing links between dataset and service metadata description, only one-way navigations) and to filter, sorting and selection functions (e.g. missing scope restriction functions, inconvenient arranged elements). These usability problems recur in various geoportals and significantly decrease the acceptance of geoportals. Therefore GI-usability patterns are suggested as a promising concept, to first summarize and categorize typical geoinformation (GI) usability problems, and second to define common solution approaches, partly being adapted from best practices in other application domains.

2 Usability Patterns and their Relations to GI-Applications

Design patterns are a well-accepted concept in software engineering [6]. As user interface (UI) design has becoming key for the acceptance of software solutions, several usability patterns, as a specific sub-set of design patterns have been suggested. These patterns should be used to improve the usability of a software product and

Table 1: Usability pattern "Direct Validation" [2].

Usability pattern	**Direct validation**
Description	When users enter data in a form that requires a specific format or has constraints on the inputs they want to identify and correct invalid entries immediately.
Solution	Validate input values during input automatically. … Show directly whether inputs are valid or invalid. Use an easy to understand, but restrained manner of presentation. Users should not be distracted and their input should not be interrupted. In case of invalid values, show the user a hint to explain the validation criteria and to correct the mistake.
Example	Creating a Google account The user must enter his current email address when creating the Google account. If the user changes to the next input field after entering the e-mail address, the system automatically validates the entered address. In case of invalid input values the system shows a specific hint (e.g. "do not forget the @-symbol")
Context	Situations in which inexperienced users need guidance for entering data Dialogs that require several input values that should be validated Free-text entries in formats that are unfamiliar or complex for users …
Rationale	Direct validation helps the users in a simple and understandable way to enter valid data. Users identify erroneous entries immediately and can correct them quickly. With specific advice on what input is expected, the system will be more conducive to learning. Time lags between data input, feedback and correction are minimized, because of immediate system responses. This avoids a change of context: Users recognize potential mistakes immediately and not after several further steps.
Conse-quences / costs	Validation requires time. Therefore validation of input values can lead to noticeable undesirable delays, which could harm users' satisfaction. In this case, a validation steps could be aggregated executed at the end of a user interaction…
Related patterns	Complement: indulgent format Complement: Auto complete

to illustrate functional solutions for usability problems in specific usage contexts, being either related to specific UI elements or to UI interaction concepts or to both [2].

Usability patterns are described by a set of design pattern attributes (name, problem description, solution) as given in table 1. However, as they stand these attributes do not provide any assistance on how to best place UI elements, or on how to best realize relations between UI elements, or towards creating consistent user interaction concepts.

Current usability patterns do clearly also describe usability aspects, which are relevant for implementing geoportals (e.g. auto complete, indulgent format or time and place-aware filters). However, they lack a specific focus on GI-applications, such as dealing with geodata types, relations between metadata and geodata or web map functions.

Consequently a GI-extended usability pattern concept is proposed, which builds on existing well-recognized patterns but also addresses geoinformation aspects. As a first subset of such GI-usability patterns, this submission focusses on geoportal implementations.

3 Usability Patterns for Geoportals

Taking geoinformation discovery as the overarching concept of a typical geoportal implementation the developed GI-usability patterns have been organized along a hypothetical discovery workflow. Thus, the patterns address the various sub-steps in such a workflow: formulating a search query, filtering results, visualising result sets on a map, etc. One important function in geoinformation discovery is map visualization. Nevertheless, in some applications it is either not implemented or easily navigating and interchanging search and map display is complicated or impossible. Here, table 2 provides an example of the general description of the GI-usability pattern *Provide map link from dataset*. This pattern tackles the issue of

Table 2: GI-usability pattern example "Provide map link from dataset view".
One pattern can be related to several attribute values (e.g. Search phases: 2 Discover of results, 3 Evaluate a result; most relevant value is underlined). Context attribute values and pattern relation types cannot be defined freely (fixed values written in italic). Attributes that do not suit a certain pattern context do not need to be set (marked with *).

GI-Usability pattern	**Provide map link from dataset view**		
Description	Users often evaluate the fitness for use of data by examine their metadata and visualization. An interactive map helps to navigate through the data. Generally, the navigation to the map is complicated (via service metadata) and needs GDI knowledge. Further, users do not know the difference between dataset and service (novice users) or need a short navigation path to the map (expert users).		
Solution	The application should provide a direct link from metadata (service as well as dataset) descriptions to the related map visualization.		
Rationale	A map serves as expressive instrument to visualize geodata. It helps users to analyse geodata visually and to evaluate the fitness for use. An interactive map can further be used to analyse the visualized data on several levels of details and to focus different regions. The map with the desired data should be easily accessible for the user. Therefore direct links to the map client are very important.		
Consequences	Map visualization should only be provided, if the geodata can be visualized on a map (e.g. standardized format). Direct links from dataset detail descriptions to the map can be realized as parameterized calls. Providing this function is more expensive than providing direct links from service descriptions, because the relation information is stored in service metadata and not in dataset metadata.		
Related patterns	Provide link from dataset view *is specialization of* Provide link to map visualization Provide link from dataset view *is similar interaction as* Provide link from service view		
Context	Activity context	Search phase	*Discover results* *Evaluate a result*
		Search dimension	*Content: Spatial extent, Temporal extent, Thematic categorization* *Result: ** *Relation: Dataset-Service* *Task: View map visualization*
		Search strategies	*Explorative search*
	UI context	UI elements	*Type: Control* *Relation: Above, Under, Next to Detail descriptions*
	User context	User types	*Novice users, Expert users*

most geoportals to (1) force users to first find a dataset and a related service description before they can view the geodata visualization and to (2) not support an easy tow-way navigation from dataset descriptions to the related map visualizations (and back), thus hampering users in executing a first visual data inspection (Table 2, Description, Rationale). As links to visualizations are only provided in web service descriptions, but are hardly given in any geodataset description, the implementation is more cost-intensive than providing map links from service descriptions (Consequences).

The pattern requires, that the map visualization (Search dimension: task) should be provided for the evaluation of a search result (Search phase) and the direct linking allows novice users to navigate to the map more easily and experts to explore the data more quickly (Description, Search strategy, User types). Regarding the UI and interaction concept, the map link has to be implemented as a UI control element, e.g. link or button, which should be placed near the dataset descriptions (UI elements) and provide the same interaction as the map link from service descriptions (related patterns).

The usability patterns for geoportals, such as *Provide map link form dataset view*, have been structured and attributed in such a way, that software designers get best supported in finding an appropriate pattern for their geoportal implementations. Table 3 provides an overview on the attributes being used to categorize the GI-usability patterns and shows the attribute domains. Thus, the attribute *search phase* allows developers to filter patterns that suit for a certain search related software part such as the provision of a result list or to generate phase-specific checklists for usability tests. Offering such search dimensions are a proven concept to classify search results and to distinguish results that match the same search term [4]. They allow developers multi-dimensional filtering and offer several entry points to discover the GI-usability pattern matrix.

GI-usability can also be organised in relation to the problems they address or according to the relations between the patterns. Figure 1 shows five map visualization patterns and their relations being classified into four types: Patterns can complement each other, e.g. "Visualize on a map" and "Provide link to map visualization", they can be dependent on each other, e.g. "Provide link to map visualization" and "Provide link back to previous page" or be related in a hierarchical structure, in which one pattern serves as specialization of another pattern. Existing usability pattern concepts do not ensure an overall consistent interaction design of an application. Therefore a pattern relation type "*is*

Table 3: Structure of Context Attribute for GI-usability patterns (attributes and sub-attributes written in bold, allowed attribute values written in italic).

Activity context	**Search phase**	*1 Formulate search query* *2 Discover results* *3 Evaluate a result* *3.1 Visualize result data* *4.1 Formulate new search query* *4.2 Filter or refine results* *5 Use results*
	Search dimension	**Content***: Spatial extent, Temporal extent, Thematic categorization Context (e.g. Organization)* **Result***: Dataset, Service, Documentation, Metadata* **Relation***: Dataset-Dataset, Dataset-Service, Dataset-Documentation, Dataset-Producer* **Task***: View map visualization*
	Search strategies	*Explorative search* *Known item search*
UI context	**UI elements**	**Type***: Input, Control, Information, Personalization* **Relation***: Above, Under, Next to, Replace by*
User context	**User types**	*Novice users, Expert users*

Figure 1: GI-Usability patterns for map visualization.

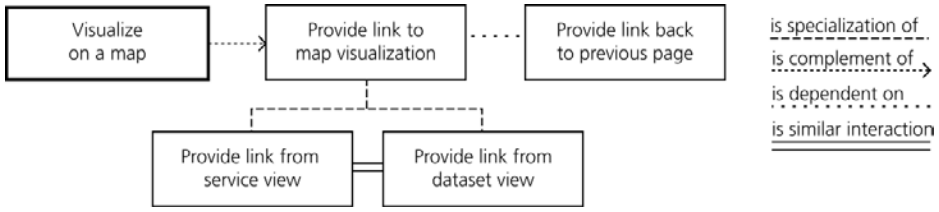
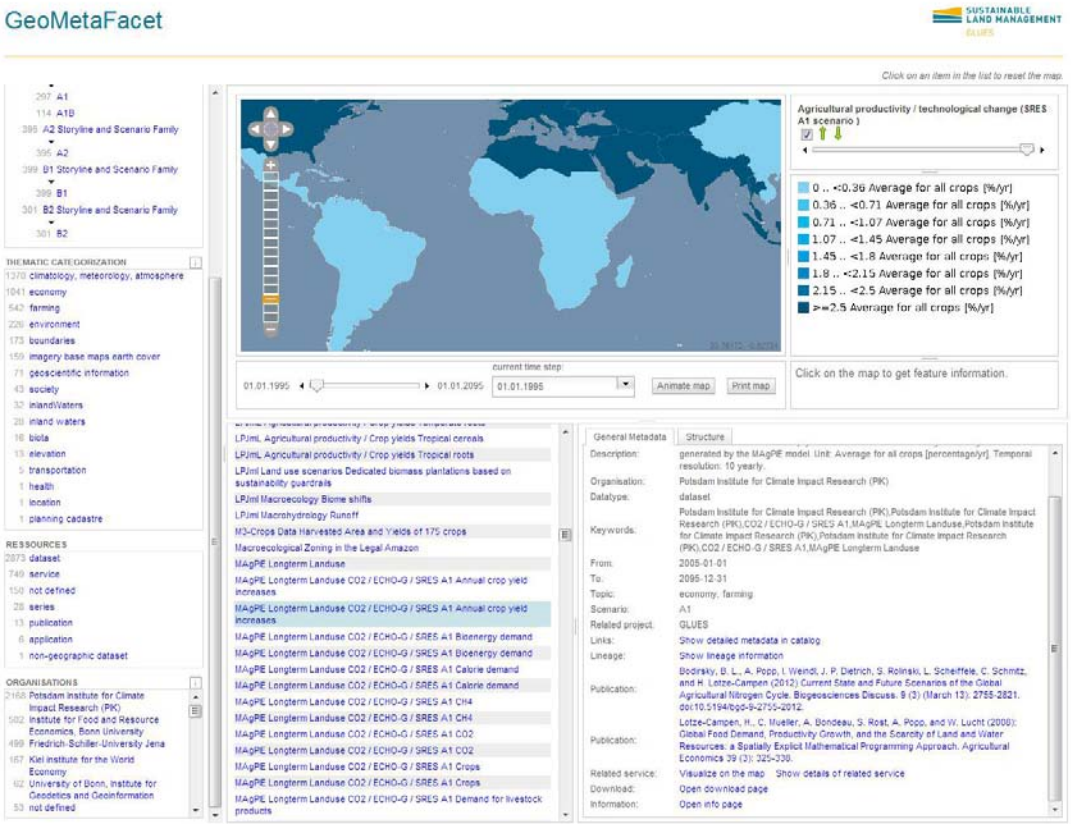


Figure 2: Reference implementation of "Provide map link from dataset view".



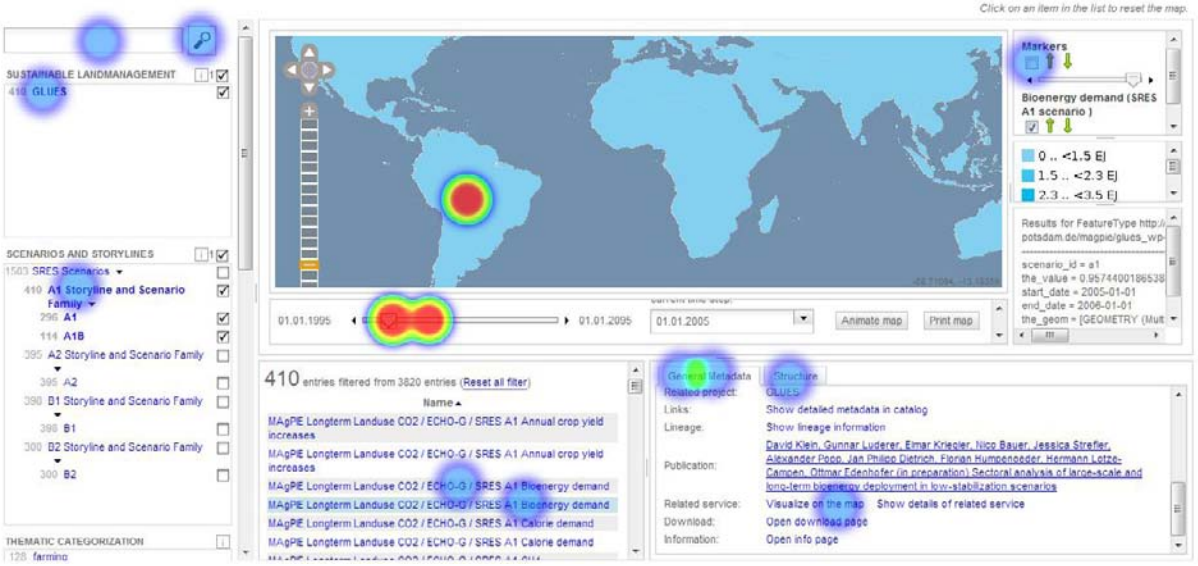
similar interaction" is proposed, to help developers in identifying all patterns which support a particular UI interaction concept.

4 Future Work

The design of GI-usability patterns is laid out as an incremental and iterative process. Having a general concept for GI-usability and related attributes a first set of patterns has been defined. This now builds the basis for further improvement of the pattern structure and thereon the definition of new pattern sets.

The defined GI-usability patterns get prototyped and exemplified in GeoMetaFacet [3], a web-client for the exploration and visualization of geodata (figure 2). This allows for future usability tests (figure 3) to help in establishing measurements for the success and efficiency of the proposed patterns. Therefore future work will investigate into qualitative (e.g. provided by ISO 9241) and quantitative metrics (e.g. eye-tracking or mouse click analysis) for these usability tests.

Figure 3: Heatmap visualization of user interactions – circles visualize mouse clicks (blue – clicked once, red – most frequently clicked).



Detailed descriptions of the patterns introduced here can be found at:

http://geoportal.glues.geo.tu-dresden.de/giusabilitypattern/index.html

References

[1] Alexander, C.; Ishikawa, S.; Silverstein, M.; Jacobson, M.;Fiksdahl-King, I.; Angel, S. (1977): A Pattern Language. Oxford University Press, New York.

[2] Röder, H. (2012): Usability Patterns, Eine Technik zur Spezifikation funktionaler Usability-Merkmale. Phd thesis.

[3] Henzen, C.; Kadner, D. (2013): GeoMetaFacet – Ein Facetten-Browser für geographische Metadaten. Geoinformatik 2013, Heidelberg (Germany).

[4] Wilson, M. L. (2012): Search User Interface Design. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, ISBN: 9781608456895.

[5] Henzen, C.; Mäs, S.; Bernard, L. (2013): Provenance Information in Geodata Infrastructures. Vandenbroucke, Danny (Ed.); Bucher, Bénédicte (Ed.); Crompvoets, Joep (Ed.), Geographic Information Science at the Heart of Europe, 2013. Lecture Notes in Geoinformation and Cartography. p. 133–151.

[6] Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J. (1994): Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley, ISBN: 0-201-63361-2.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Agile access to sensor network

Sergio Trilles Oliver
INIT
Universitat Jaume I
Castelló de la Plana, Spain
strilles@uji.es

Óscar Belmonte Fernández
INIT
Universitat Jaume I
Castelló de la Plana, Spain
belfern@uji.es

Laura Díaz Sánchez
INIT
Universitat Jaume I
Castelló de la Plana, Spain
diazl@uji.es

Joaquín Huerta Guijarro
INIT
Universitat Jaume I
Castelló de la Plana, Spain
huerta@uji.es

Abstract

The work in this paper aims at increasing the interoperability and improving accessibility of data provided by sensor networks. This way, this data can be employed by different devices and with diverse context requirements, such as specific location and time. To address this problem Geographic Information System (GIS) services, such as the Sensor Observation Service (SOS), in conjunction with Representational State Transfer (RESTFul) architecture are used. A standard-based solution that increases interoperability is presented. It also allows for a better integration of data already published in different semi-structured formats in order to be used by various platforms (web or mobile). Furthermore, this system adds value to original sensor data in order to assist in the decision making process.

Keywords: air quality sensors; meteorological sensors; heterogeneous sensor sources; RESTFul services; sensor observation services

1 Introduction

This paper describes a system for the processing of sensor data, the publication of these data as interoperable services and their interoperable access from multiple platforms. The goal of this work is to provide interoperable access to heterogeneous data formats coming from environmental sensor networks by using standards or agile interfaces for enabling its easy access. To achieve this, Open Geospatial Consortium (OGC) standards following the INSPIRE architecture [1] are used. INSPIRE should provide environmental data related to 34 themes, including transport networks, land cover and hydrography, INSPIRE provides important parts of the European contribution to a Global Earth Observation System of Systems (GEOSS).

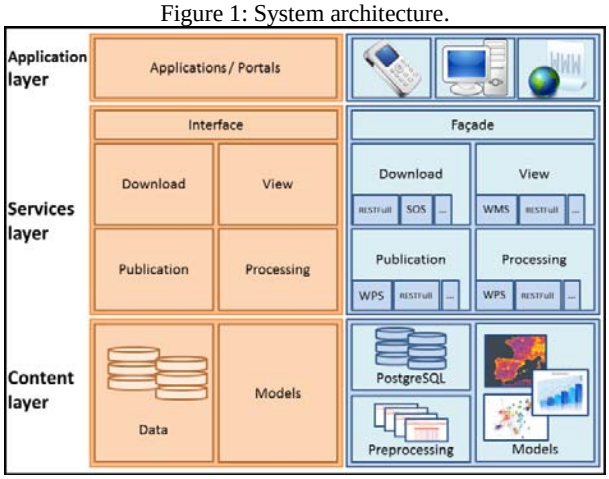
Standard format such as the Sensor Observation Service (SOS) are used. SOS provides a standardized web service interface that allows clients to access descriptions of associated sensors and their collected observations [2]. Other types of interfaces are provided which are capable of lightening the use of these data sources to be consumed by more restrictive devices such as mobile phones, although without compromising the interoperability standards. In this project an interface Representational State Transfer (RESTFul) has been used, following the principles of Internet of Things (IoT) [3].

Finally, this work is completed with an analysis models. A hotspot model is designed following Air Quality Index (AQI). For this model the OGC standard is used to process spatial data, Web Processing Service (WPS) [4]. WPS specification is generally defined to provide standard-based access to a wide variety of calculations of spatial data using web services. The use of WPS ensures service interoperability.

2 System design

The present system follows the INSPIRE technical architecture, which contains three layers [2]: content layer, service layer and application layer. At the top layer reside the users and applications (clients). At the middle layer reside all services that provide the required functionality such as accessibility and processing of data. The INSPIRE initiative guarantees interoperability in such systems.

Figure 1: System architecture.



This system operates in the three layers of the Spatial Data Infrastructure (SDI) architecture. In the content layer, it must perform a preprocessing to integrate different and distributed sources into this single system. In the services layers, the different sources are interrelated, (geo) visualizing and processing data to derive useful information by generating

indexes or predictions such as pollutant concentrations in an area. In addition, our system should allow access to these services from any device that has Internet access, from a desktop, web or mobile application. In the application layer, client applications consume services provided by the previous layer.

Figure 1 shows the conceptual architecture of this system. On the left side, the conceptual view the system's architecture is shown. In the content layer, the data is found, obtained from different heterogeneous sensor sources. In addition scientific models are also found that they use to model and process the data. In the services layer, four different services are found as well as an extension of the functionality as recommended by INSPIRE; download service, view service, publish service and processing service. Finally, in the application layer, applications and portals for displaying information can be found.

On the right side of the figure, there is a more technological view of the proposed system. At the bottom, is the pre-processing module, where the system implements the preparation and integration of the information in different structures, such as databases.

Once the data has been processed, it is inserted into the database. The PostgreSQL databases are used. In this way, multiple data sources are integrated with different structures and characters, and are offered in a structured and interoperable way.

Different scientific models are defined, which are applied to process the data and derive required information: for example, view (heat maps), analysis (clustering), prediction or propagation.

The different services are deployed: data and processing Services, where published data and processes based on standard interfaces, such as, SOS, WPS [4], Web Map Service (WMS) [5] and other light-weighted interfaces, such as, RESTFul to implement data and map services. In order to develop a scalable and extensible system a pattern design is followed [6], to develop and expose components. The Façade pattern is used. It will be able to offer a single entry point for all services. The multiple interfaces are available through a façade component able to encapsulate the complexity of various interfaces to obtain entry to the same data to increase interoperability. The aim is to provide a variety of interfaces, without losing interoperability of GIS standards, to publish and consume sensorial data. This allows us to increase the variety of information received.

A RESTFul service is offered, to download sensorial data. This allows for a better management of the SOS interface [7]. In this way, a RESTFul interface is developed that encapsulates the downloading service.

Furthermore, the processing services are also deployed offering different interfaces for the same functionality in order to increase interoperability and better adaptation to the client applications.

The right side of the figure shows the interaction with the user via the client applications. In the first approach, and due to the increasing demand in mobile devices, a client application is developed and these services are provided with lighter interfaces for efficient communications.

3 Services

This section details the different services created for developing this system. As already explained, a RESTFul interface is created, this interface is adapted to the SWE services. Also, a SOS service and another service that returns the KML's are built. Finally, hotspot model is shown.

3.1 RESTFul services

As a first approximation to the accessing and downloading service, a service that implements a RESTFul interface is designed. A RESTFul interface is an architectural model that can be efficiently implemented as a combination of the Hypertext Transfer Protocol (HTTP) and TCP/IP. In addition, the RESTFul interface offers better access from any device, even from mobile phones [8], which have inferior technical features.

The RESTFul web service has been developed following the paradigm of IoT and more specifically Web of Things (WoT) .WoT is the evolution of IoT [9], which uses web standards such as RESTFul for implementation. In the concept of the WoT, web servers are embedded into everyday objects, where they were turned into "smart things" [10].

The RESTful enables access to sensor resources by Uniform Resource Identifier (URI). The service supports the listing of all sensors and the retrieval of their sensor data. Connected sensors are listed at the entry point of the sensor platform. The output format is JavaScript Object Notation (JSON). JavaScript Object Notation (JSON) has the advantage of being more compact than eXtensible Markup Language (XML). JSON is a lightweight text-based data exchange format; read- and writable by humans as well as parse and generate by machines [11].

Sensor data is returned when accessing the corresponding resource. The pattern of the URI-Scheme is defined as:

http ://<server>/<sensorId>/<method>

<server> the entry point of the sensor platform. It provides a collection of sensors which are attached to the platform.
<sensorId> refers to an identifier for a specific sensor. When accessed this resource should list a collection of all available methods applicable for this sensor.
<method> stands for the method of interaction with the sensor.

3.2 Sensor Observation Service

This system also provides an SOS interface which guarantees a standard interface capable of using sensor data in an interoperable way. The implementation of 52North (SOS version 2.0) is used. SOS has three core operations which provide its basic functionality: GetCapabilities, DescribeSensor and GetObservation.

3.3 Vector Sensor data format

Another RESTFul operation is designed that is capable of exporting the data in a Keyhole Markup Language (KML) file. This operation has 3 parameters, "date1" with the initial

date, "date2" with the deadline and "comp" component to show.

The generated KML shows columns graphics for each sensor, and it shows a different color for the three defined levels: good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy and Hazardous (Table 1). These levels are defined following AQI, only for the components: O, NO2, SO2, CO and PMx.

3.4 Hotspot model

A model is created that classifies observation through AQI. This processing service is designed that is to be implemented as a standard-based processing service. This service offers one operation, which through air quality observation, classifies this value on one level (Table 1).

Table 1: Different levels of thresholds.

	AQ Index
Good	0-50
Moderate	51-100
Unhealthy for Sensitive Groups	101-150
Unhealthy	151-200
Very unhealthy	201-300

4 Use case: heterogeneous air quality and meteorological data sources

Two different data source are worked with in this process. One of them is the Valencia network of surveillance and control of air pollution[1], which is able to perform analysis of the air in real-time. Although in this article we focus on air quality in the Valencia network; we have also worked on data from the Meteoclimatic network.

Figure 2: **(a)** List of air quality stations. (**b**) last observation in map.

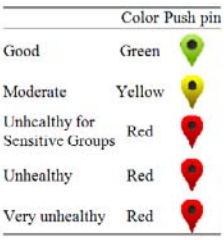
For a first prototype, an application for the Windows Phone OS has been created. Figure 2(a) shows a screen shoots of the mobile application where the list of all sensors and users can view information about station and historical observation. Figure 2(b) shows how the sensors are placed using a pushpin and user can see last observation about active pushpin.

The client offers the possibility to help to make decisions. The map view mode allows visualization of hotspots. A hotspot is considered when any measure exceeds a permissible level. So the thresholds that AQ Index establishes are taken. In the following table (Figure 3(a)) different colours for each contaminant level have been set.

In this way, when a component of a station has a superior level that its threshold, the pushpin is shown in the colour that it has defined in the table (Figure 3(b)). When you click on the pushpin, the colour of the component value appears in the colour that corresponds to the measured level.

These points can help to make a decision, for example: route planner. This way, it is possible to obtain a route that is far from any hotspot.

Figure 3: **(a)** different levels of thresholds and. (**b**) map with hotspot



a)

b)

5 Conclusions

At present, there are many data sources about sensors, but they have a problem, they contain heterogeneous data. An important challenge is to increase the interoperability of this data sources. In this paper a service-oriented architecture is proposed to implement an application, which orchestrates a workflow for the management of heterogeneous data formats provided by sensor network. A system that aims at the integration of unstructured data sources with different character is proposed, and which offers them in a structured and interoperable way. In addition, the service layer in our application is enhanced with a module that implements the design pattern façade, and acts as a "middleware" between the client and the services to increase interoperability by implementing several interfaces and helping users to get information easier.

This is a recent and ongoing research project and there are still questions to be answered and tasks to be implemented in the future, such as, the addition of new sources of information, the development of the middleware (façade), whether to increase the interoperability of the services, the addition of the processing models to communicate more sophisticated information to users, such as display of not only measured

[1] http://www.cma.gva.es/web/indice.aspx?nodo=4581&idioma=C

sensor values but also simulated values to run news propagation and prediction models...

References

[1] EU Directive. Directive 2007/2/ec of the european parliament and of the council of 14 march 2007 establishing an infrastructure for spatial information in the european community (inspire). Technical report, Official Journal of the European Union, L 108/1, Volume 50, 2007.

[2] A. Na, M. Priest. OGC Sensor Observation Service Implementation Specification, OGC 06-009r6, Open Geospatial Consortium, Inc.

[3] S. Uckelmann, M. Harrison, F. Michahelles. An Architectural Approach Towards the Future Internet of Things; Uckelmann, D., Harrison, M., Michahelles, F., Eds.; *Architecting the Internet of Things*. Springer Berlin Heidelberg, 2011

[4] P. Schut. Opengis web processing service version 1.0.0. OpenGeospatial Consortium., 2008.

[5] J. De La Beaujardiere. OpenGIS Web Map Service (WMS) Implementation Specification. http://portal.opengeospatial.org /files/?artifact_id=5316 (accessed on 15 January 2014)

[6] E. Gamma, R. Helm, R. Johnson. Vlissides, J. Dessign Patterns: Elements of Reusable Object-Oriented Software, Addison Wesley, 1995.

[7] M. Rouached, S. Baccar, M. Abid. RESTful Sensor Web Enablement Services for Wireless Sensor Networks. *IEEE Eighth World Congress* 2012, 72, 24-29.

[8] H. Hamad, M. Saad, R. Abed. Performance Evaluation of RESTful Web Services for Mobile Devices. *International Arab Journal of e-Technology* 2010, 1.

[9] S. Duquennoy, G. Grimaud, J. Vandewalle. The Web of Things: Interconnecting Devices with High Usability and Performance. In *Proccidings of the International Conference on Embedded Software and Systems, ICESS '09*, pp. 323–330, Hangzhou, Zhejiang, China, 2009.

[10] D. Guinard, V. Trifa, F. Mattern, E. Wilde. From the Internet of Things to the Web of Things: Resource Oriented Architecture and Best Practices. Version: April 2011. http://www.vs.inf.ethz.ch/res/papers/dguinard-fromth-2010.pdf (accessed on 15 January 2014).

[11] D. Crockford. The application/json Media Type for JavaScript Object Notation (JSON). RFC 4627 (Informational). http://www.ietf.org/rfc/rfc4627.txt. (accessed on 15 January 2014).

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Utilization of NoSQL database for disaster preparedness

Winhard Tampubolon
AGIS, Institute for Applied Computer Science
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39, 85577
Neubiberg, Munich
winhard.tampubolon@unibw.de

Abstract

Nowadays, in the age of big data, geodatabases become more critical with respect to geospatial data volume, variety and capacity. It is required that geodatabases must be capable enough to cope with high stakes of geospatial data service during production, manipulation and publication stages.

The concept of NoSQL database has been introduced as a potential alternative solution to existing SQL databases which is supposed to grow more rapidly in the near future. It has the prospective to combine the powerful capability of GIS data processing with an approach of non-relational Data Base Management System (DBMS). This type of data warehouse can potentially accommodate variety of information over the World Wide Web (www) space with different structures into one single geodatabase. MongoDB as one instance of NoSQL database introduces an open source document storage empowered by a replication using data partitioning approach across multiple machines.

For the work described in this paper it has been used for the integration of open access geo-information by extracting geospatial information from a near real time earthquake service i.e. Geofon. Geospatial information is extracted from the Geofon uniform resource locator (url) then transferred into documents in MongoDB. This demonstrates the geospatial data integration in order to improve earthquake information contents as well as to enable GIS analysis approach using Python scripting environment in ArcGIS 10 platform. It shows a reliable performance even for handling a relatively big geographical names data from GEOnet Names Service (GNS).

Keywords: Earthquake, NoSQL, non-relational, rapid mapping, geographical names

1 Introduction

The occurrences of disasters all over the world trigger the awareness of many responsible institutions around the globe to deliver useful services in order to perform disaster preparedness.

Using a Desktop GIS as a basic processing platform, this practical work has combined two point features from different sources into one single geodatabase in MongoDB.

The motivation of this paper is to integrate near real time information from web service in a GIS Desktop using NoSQL database (MongoDB) as the data warehouse by enabling geospatial functionality.

2 NoSQL Database

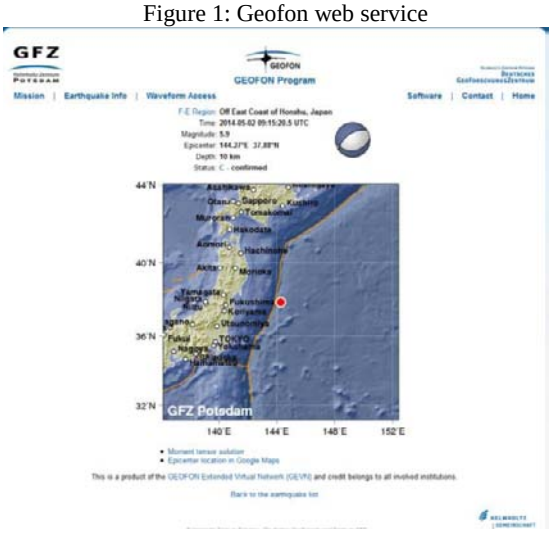
NoSQL databases introduce a new approach to overcome the problem of data structure inconsistency as a challenge to the conventional Relational Data Base Management System (RDBMS) [2]. The most important factors that trigger NoSQL technologies were the uprising of crowdsourcing and technology driven demand [1].

MongodDB as an instance of NoSQL database implements schema flexibility by using Java Script Object Notion (JSON) format and document based approach [3]. Actually those two approaches will combine the advantages of flexible data structure and data transfer capabilities.

3 Earthquake webservice

Geofon is one instance of the earthquake information service operated by German Geo-research Centre (GFZ) which is basically in near real time manner. Although the sensor platform is always switched on to monitor seismic activities but usually it also analyses recorded data before any further public distribution (Figure 1).

Figure 1: Geofon web service



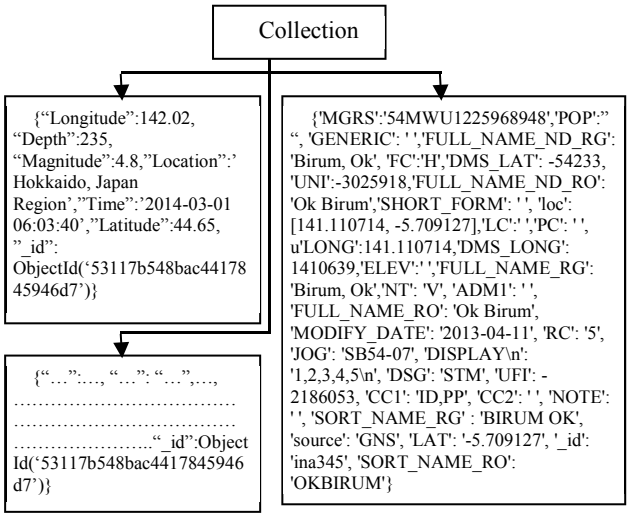
4 Integration concept

Data integration has been performed by the following technical implementation steps:

1. Geospatial data creation

   It combines capability to read, filter and transfer geospatial content from Geofon and GNS into MongoDB, in the context of disaster preparedness. Those two features have been stored in one single data MongoDB collection on a document basis (Figure 2).

Figure 2: Document based storage



2. Real time integration

   By simply implementing scheduled task, a python script will be executed on each frequent time i.e. every 5 minutes in order to get real access to seismological monitoring network around the globe maintained by Geofon service.

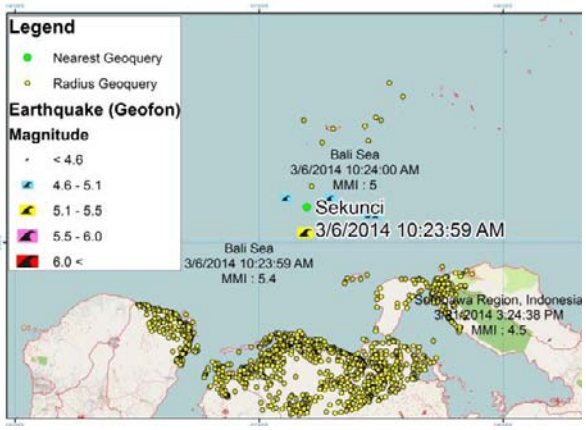
3. Geospatial Index and Query

   By using subsequent updating for the whole document, it creates one additional field i.e. 'loc' that contains Longitude and Latitude data in a float values for geospatial index purposes.

   As an instance to improve the "location" field in Geofon earthquake service, the "near" function in MongoDB has been implemented to get the nearest feature from latest earthquake occurrence. It improves the location field from "Bali Sea" to more accurate result of "Sekunci" (name of the island) automatically (Figure 3).

   In addition using "within" radius functionality, field information from different features can be joined as a complementary intersection in geo-processing. This capability has been tested to join field within a certain radius which yield ArcGIS layer with additional field from earthquake event e.g. Magnitude, Depth and Time.

Figure 3: Geospatial functionalities



5 Conclusions

Geospatial information can be very useful if it is integrated with other information system. This practical work has integrated real time earthquake data from Geofon with GNS. As the final result, it brings real time information improved by static geospatial data source from GNS.

The NoSQL database plays some important roles in the integration system both as geospatial data storage and geospatial data processor. Geo-processing performance in MongoDB has been tested with reliable geo-querying function such as 'near' and 'within' in ArcGIS 10.

Real time integration is mainly supported by scripting capability in simple non-SQL query which enables geospatial analysis from different features and structures using NoSQL database approach.

References

[1] Harrison, J. Using NoSQL & HTML5 Libraries to rapidly generate interactive web visualizations of high-volume spatiotemporal data, Association for Geographic Information (AGI), London, UK, (online last accessed on 03.05.2014 http://www.agi.org.uk/storage/GeoCommunity/AGI13/Speakers/ppts/JackHarrison.pdf), 2013.

[2] COUCHBASE, Making the Shift from Relational to NoSQL, Whitepaper, (online last accessed on 03.05.2014 http://www.couchbase.com/sites/default/files/uploads/all/whitepapers/Couchbase_Whitepaper_Transitioning_Relational_to_NoSQL.pdf), 2014

[3] MONGODB, Making the Shift from Relational to NoSQL, Whitepaper, (online last accessed on 03.05.2014 http://info.mongodb.com/rs/mongodb/images/10gen_Top_5_NoSQL_Considerations.pdf), 2013

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Latest Developments and activities in the Spanish NSDI

Antonio Rodríguez
Instituto Geográfico Nacional
Spain

Abstract

The Spanish NSDI Geoportal was opened on June 2004 and has evolving continuously with the support of a healthy collaborative community composed by public and private sector, academia and citizens. INSPIRE Directive has been transposed in Spain by means of Law 14/2010, LISIGE, the Law of Spanish Geographic Information Services and Data, which establishes a national coordination structure based on the High Geographic Council as umbrella organization embracing the Spanish Geographic Information Infrastructure Managing Board and a set of Technical Working Groups (TWGS). The TWG on Monitoring and Reporting has been coordinating this task since 2010. In this communication a brief summary of the state of the play of the project is provided, including achievements, conclusions, lessons learnt and good practices, giving special attention to and the process of monitoring and reporting.

Keywords: SDI, web services, INSPIRE, monitoring, reporting.

1 Introduction

INSPIRE Directive prescribes that Member states should provide descriptions of spatial datasets and services within the scope of the National SDIs and shall establish and operate a catalogue service for the spatial data sets and services for which metadata has been created.
To ensure the interoperability of data and services in the European SDI, a set of Implementing Rules (IR) have been adopted for all the relevant aspects implied: metadata, network services, spatial data services, monitoring and reporting and data and services sharing.
Taking into account the principles of INSPIRE Directive about services and looking for compliance with Implementing Rules, IGN Spain initiated in 2011 a reengineering process of its more basic services: a viewing service (IGN-Base) and a discovery service (IGN-CSW).
The adaptation process is still running in a continuous adaptation approach. New analysis, design, implementation and tests are performed to put in practice the legal and technical framework defined by INSPIRE Directive and its Regulations.

2 Inspire web services at IDEE

To implement INSPIRE Philosophy and principles in the national node of Spanish NSDI, a considerable efforts has been made during 2012 to establish a core set of INSPIRE compliant and useful harmonized services integrated by:

- IGN visualization services, a set of WMS and WMTS INSPIRE services conformant Technical Guidance for the implementation of INSPIRE View Services. Those services have been made with a free software (Geoserver 2.2.4 with Inspire extension) with some tuning and additions.
- IGN-CSW, a INSPIRE discovery service (CSW) that allows the searching and retrieval of descriptive information (metadata) about data and services, and meeting the Technical Guidance for the implementation of INSPIRE Discovery Services requirements. This service has been also implemented with a free software application (Geonetwork 2.6.4.).
- IGN download services, including predefined dataset download services (ATOM) using new developments and direct access download services (WFS 2.0) using a free software (Geoserver 2.2.4 with app-schema extension). Those services fulfills Technical Guidance for the implementation of INSPIRE Download Services requirements.

All those services have been described with metadata and they are available at the INSPIRE Geoportal through IDEE CSW service and any user can access and consult them. In addition, IDEE CSW provides access to more than 80,000 metadata registers of services and spatial data of IDEE community. A procedure of coordination among the network of 13 discovery services has been implemented based on harvesting and metadata XML files interchange.

3 Inspire web services at IDEE

By the other hand, a big effort to analyze, design and develop software tools to validate and assess conformance and quality of the resources implemented in the considerable set of SDI nodes existing in Spanish NSDI is being done.
Under a collaboration agreement with University of Zaragoza, IGN Spain has begun to develop the following utilities:

- An on-line Inspire Services Validator (ISV) for INSPIRE Network Services to validate its conformance with the applicable regulations and standards. In a first phase visualization services and discovery services have been considered.

- An on-line Inspire Metadata Validator (IMV) for testing INSPIRE conformance of metadata.
- A new optimized version of the Open Source metadata editor CatMDEdit taking into account all the INSPIRE relevant requirements for the generation of data and services metadata.

Finally, in relation to the assistance given by the national node for the implementation of Inspire in Spain, a new INSPIRE special section has been created in the IDEE Geoportal, where users can view and access documents and recent developments that can help them to the correct implementation of INSPIRE Directive in its organization.

4 Monitoring and reporting

On the other hand Article No. 21 of INSPIRE Directive states that Member States should carry out monitoring of the implementation and use of their infrastructures for spatial information. This annual review should be sent to the Commission and published by 15 May each year. The first INSPIRE monitoring process was conducted in 2010.

The transposition of INSPIRE Directive to the Spanish legal framework was completed in 2010 with the approval of Law 14/2010 of the 5th of July, the Law on Infrastructure and Geographic Information Services in Spain (LISIGE). Among other things, this law creates organizational structures in the Spanish administration meant to implement and comply with the requirements defined in INSPIRE Directive and its Implementing Rules.

LISIGE assigns to the High Geographic Council (CSG) the role of coordinating the implementation of the INSPIRE Directive in Spain. This results in the creation of the Spanish Geographic Information Infrastructure Managing Board (CODIIGE). CODIIGE, which was established in 2011, has among its functions to gather information on interoperable geographic data and services, and also to establish the criteria to select the information to be sent from Spain to the EC in the annual monitoring report required by the Directive.

To support CODIIGE activity, in November 2011, 17 Technical Working Groups (TWGs) were created, composed by senior representatives of the agencies responsible for the production of geographic information in Spain. One of them is the GTT monitoring and reporting (GTT S & I), which has among its specific duties the following: coordinate and conduct the annual collection of data sets of data, metadata and services from the bodies of the Central Government, Autonomous Communities and Autonomous Cities that are required to carry out monitoring, and to coordinate and conduct the triennial gathering of information from the institutions involved in the follow up to compose the report required by INSPIRE Directive.

The first monitoring and reporting process took place in 2010, coordinated by the National Geographic Institute (IGN) as Technical Secretariat of CSG. From 2011 the monitoring process was coordinated by the TWG S & I and has taken advantage of the participation of the other experts from CODIIGE TWGs.
Public agencies, technicians and national experts have taken part in all the processes of monitoring, being asked for information. After four monitoring seasons and two tracking reports, it is observed that the number of participants tends to stabilize and the quality of the responses has been improving step by step. This is mainly due to the fact that the different coordination structures, at state and regional levels, has been assuming and improving its roles and performance with respect to INSPIRE obligations.

The gained experience has made possible the adaptation and adoption of the necessary measures to address the emerging issues and problems. Hopefully it will allow to propose future actions leading to the publication of data sets and services in fully accordance with INSPIRE data specifications and technical requirements.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

A Spatial Approach to Surveying Crime-Problematic Areas at the Street Level

Lucy Waruguru Mburu
GIScience Group
Heidelberg University
Berliner Straße 48
69120 Heidelberg,
Germany
lucy.waruguru@geog.uni-heidelberg.de

Alexander Zipf
Chair, GIScience Group
Heidelberg University
Berliner Straße 48
69120 Heidelberg,
Germany
zipf@uni-heidelberg.de

Abstract

Reaching far beyond the realm of geography and its related disciplines, spatial analysis and visualization tools now actively support the decision-making processes of law enforcement agencies. Interactive mapping of crime outperforms the previously manual and laborious querying of crime databases. Using burglary and robbery events reported in the urban city of Manchester, England, we illustrate the utility of graphical methods for interactive analysis and visualization of event data. These novel surveillance techniques provide insight into offending characteristics and changes in the offending process in ways that cannot be replicated by traditional crime investigative methods. We present a step-wise methodology for computing the intensity of aggregated crime events which can potentially accelerate law enforcers' decision making processes by mapping concentrations of crime in near real time.
Keywords: Crime, spatial visualization, kernel density estimation, decision support.

1 Introduction

Criminal events which accumulate over the geographic space have severe consequences, such as creating fear and general distrust among residents [1]. Burglaries for example are a common occurrence which deteriorates the economic framework of urban cities by discouraging local and international investors. Researchers, planners and law enforces can increase urban safety by predicting criminal events. Identifying common patterns of crime distribution across space enables the strategic placement of order enforcement mechanisms and programs.

2 Literature Review

The potential of dynamic visualization to explore temporal changes in spatial data by manipulating spatial displays is clear. Brundson et al. compare the effectiveness of three visualization techniques; map animation, comaps, and isosurfaces [2]. Later studies have also employed interactive and dynamic visualization tools to map crime patterns within the spatial and temporal contexts [3, 4]. At local, regional and global conferences discussions of the capabilities of various tools to analyze crime within many applications have become commonplace.

To understand how to effectively apply such tools, crime mappers and law enforcers use various interdisciplinary principles that demystify offender behaviour. Three such principles which underlie the methods used in this study, are time geography, routine activity and the offender's rational choice [5, 6]. Time geography explains the rhythmic variations of human activity. Constricted by movement, people converge during specific spatio-temporal windows through the transport system. This forces social interaction and creates audience between criminals and their victims [7]. While recognizing this interaction, routine activity on the other hand explains crime to result from three elements uniting; a motivated offender, an attractive and unguarded victim or property, and the absence of a guardian to prevent the crime. Reducing crime therefore requires removing offenders or increasing safety guardians. Finally rational choice explains the offender as one who weighs the risks and returns associated with each crime. Guided by this principle we compare the occurrences of burglary alongside robbery which has higher associated risk and returns, to identify the variability in spatial and temporal signatures of different offenders.

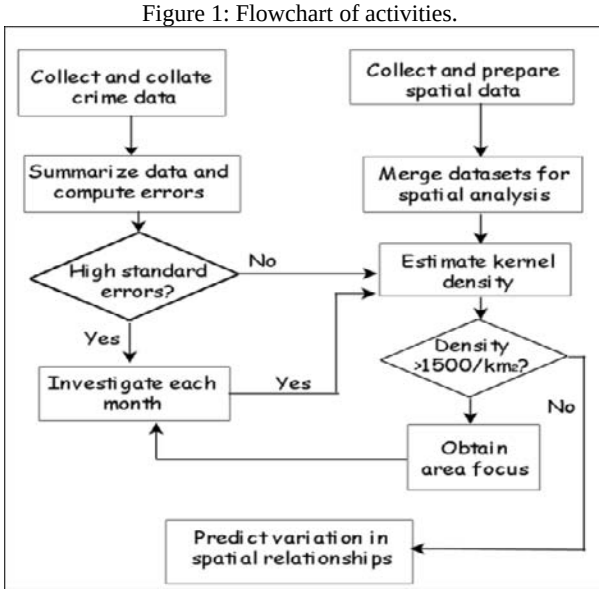
While each single criminal record contains vital and unique information, the collective mapping of multiple events in time and space unearths underlying patterns which inform the decision support processes of order enforcement practitioners. Hotspot mapping of crime events is a common and innovative use of crime mapping. To visualize the spatial and temporal dimensions of crime, several studies [8, 9] recommend computing kernel density for a locale of interest.

With the blossoming of spatial analysis and GIS availability, empirical geography of crime is now embedded within the justice system of England and Wales [10]. This paper contributes to the existing body of literature by providing a systematic approach to analyze and visualize patterns of crime disturbance in the urban city of Manchester, England.

3 Methodology

We introduce a step-wise statistical analysis and visualization process to identify general areas affected by high criminal activity and to focus on these areas. This involves the simultaneous computation and display of descriptive summaries to pinpoint problem areas, and the subsequent spatial analysis and display of spatial output (Figure 1). We primarily used two packages of the R language; the SPAtial Relative Risk (sparr, [11]) and ggplot2 [12] to analyze and visualize the concentration of crime respectively. The flexibility of R allows the reuse of code functions among datasets, graphical displays and devices.

Figure 1: Flowchart of activities.



We estimated relative clustering of crime with kernel density estimation (KDE). This exploratory technique estimates the density of aggregated point events which lie within a defined boundary. An intensity variable (z – value) estimates density for all parts of an area. Resultant measures are displayed as surface maps. Such maps provide a certain level of abstraction at which areas in need of security prioritization can be clearly identified while the private details of individual crime events remain hidden.

We summarized crime information to obtain means and standard errors. To generate risk surfaces we used an adaptive bandwidth where the kernel width is varied in different regions of the sampling space. For each observation with positional information in two columns, data[i, 1:2], i = 1,2,..n), the bandwidth, h[i] is derived by:

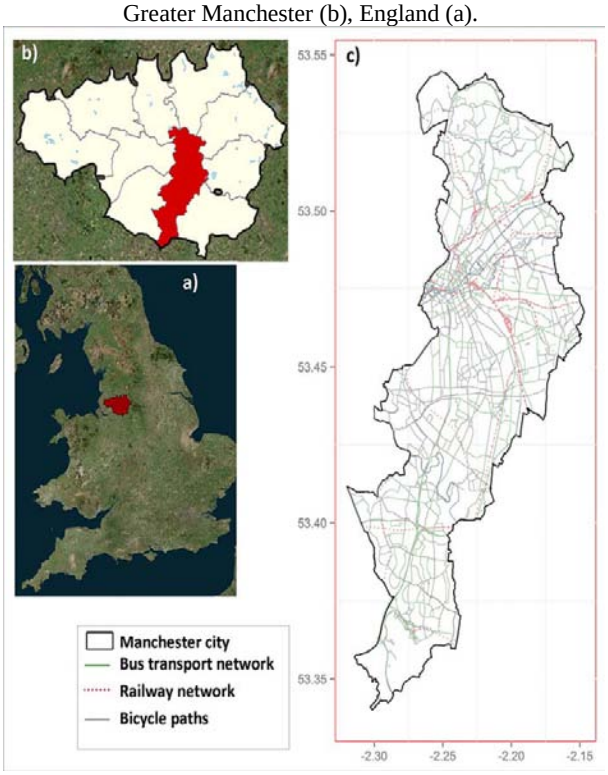
$$h[i] = globalH/(w(data[i,1{:}2];\ pilotH)\text{^}(1/2) * gamma)$$

Here w denotes the fixed bandwidth pilot density that is constructed with bandwidth *pilotH* (a positive smoothing parameter), and the scaling parameter gamma is the geometric mean of the *w^(-1/2)* values. We then identified an appropriate extent for the search of clusters.

4 Study area and Data

The city of Manchester comprises 33 wards and has a well developed road network infrastructure. Primary generators of crime are the city's nodes of transportation (Figure 2). To visualize the influence of transport nodes on offending behaviour we obtained shape files of public bus routes, the railway network, bicycle paths and the waterways within the city of Manchester. This spatial data are provided by the Great Britain's national mapping authority through the OS Openspace application programming interface (API) under the UK's Open Government License (OGL).

Figure 2: Transport network of the city of Manchester (c) in Greater Manchester (b), England (a).



We obtained from the UK Government police API 28,340 burglary and robbery events reported between 1 January, 2011 and 31 December, 2013 by the Greater Manchester police agency. Table 1 shows a summary of the observation data. These figures provide an overview of the offending patterns in Manchester. As described by the work flow, high standard errors with monthly crime aggregates necessitated investigating crime data for each month. An additional observation is the sparse distribution of crime across the temporal space (e.g. 1.3 per square m), which suggests unequal distribution of criminal activities. This observation calls for the visualization of the crime events together with their associated spatial information to observe if certain areas experience increased offending rates.

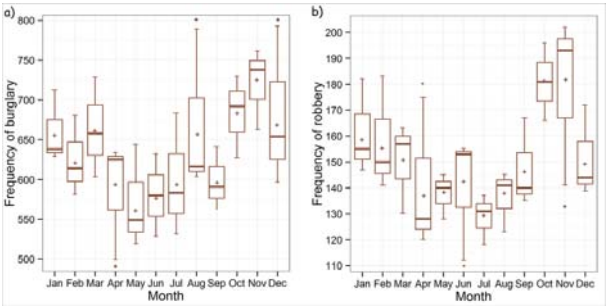
Table 1: Burglary and Robbery frequencies in the city of Manchester

Crimes	Count	Mean Annual Crime	Mean Monthly Crime	Annual crime/km2	Monthly crime/km2
Robbery	5435	1811.67 (min=1691, max=1929, σ= 119)	150.97 (min=112, max=202, σ= 23)	47	1.31
Burglary	22905	7635(min=7165, max=7946, σ=344.3)	636.25 (min=499, max=792, σ=73.6)	198	5.50
Total	28340	393.61	393.61	245	3.40

5 Results and Discussion

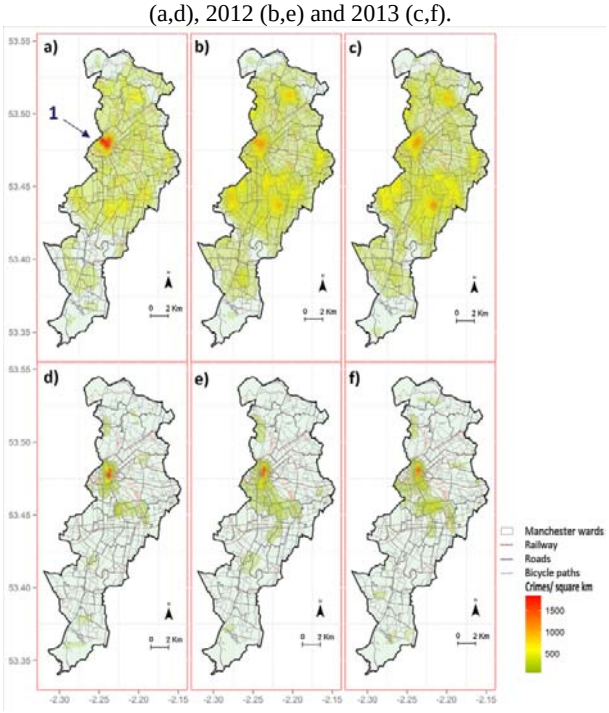
Figure 3 provides a summary of monthly crime data. Visualizing crimes in this way allows the comparison of the crime distribution. More burglaries than robberies are, for example, observed during the study period, and this can be explained by the theory of rational choice discussed above. While robberies often provide higher returns than burglaries, the risk of apprehension during robberies is often higher, as is the penalty to the apprehended robber. Both categories of crime have reduced activity between April and September, and increased activity in October and November.

Figure 3: Burglaries (a) and robberies (b) in the months of 2011 to 2013.



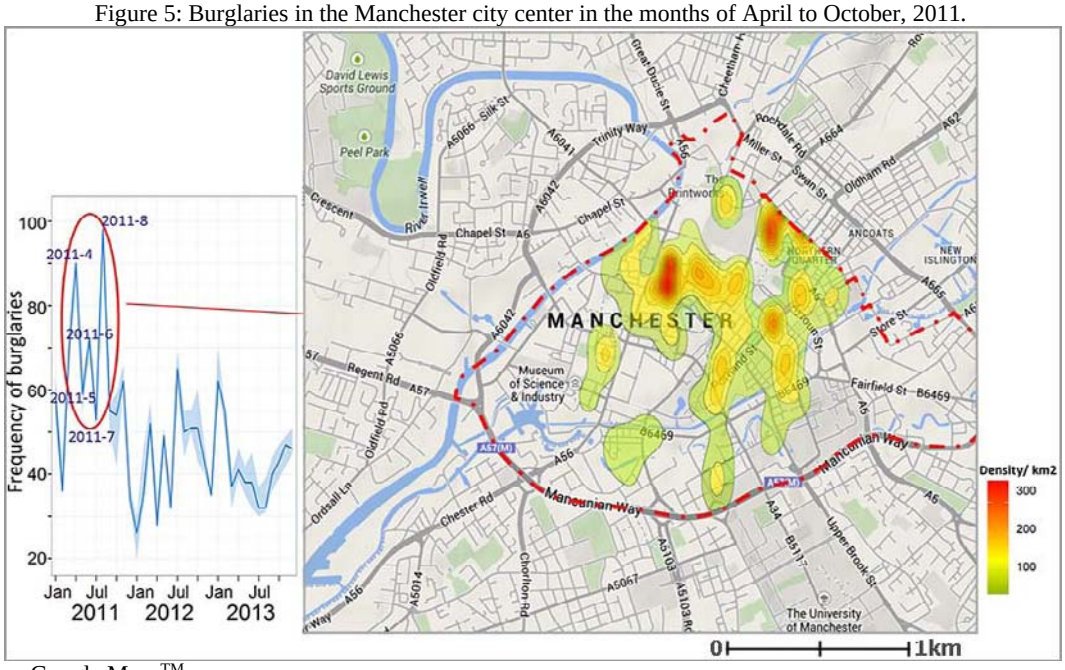
To observe the geographic distribution of crime we visualized density estimates side by side (Figure 4). The most frequent burglaries were observed in 2013 (Figure 4c). The highest concentration of burglaries in space, labeled "1", was however observed at the city center in 2011. Robbery hotspots were also consistently observed near the city center, with the widest distribution observed in 2013. The city center has the busiest network of railway, roads and bicycle paths. This results in constant human traffic, and, as explained by the theory of crime geography, it creates an opportunity for offenders and their victims to converge.

Figure 4: Concentrations of burglary and robbery in 2011 (a,d), 2012 (b,e) and 2013 (c,f).



We obtained a detailed focus of the city centre to visualize crime incidents at the street level with Google map using ggmap's [13] function, get_map() (Figure 5). The output explains the same variability in offense distribution as had been observed previously, especially with burglaries. The month of August, 2011 in which the highest number of burglaries are visualized, witnessed one of the most intense riots and random looting of property in the city centre of Manchester [14]. Such an illustration makes obvious the potential that spatial analytical and visualization tools have for measuring historical events and pinpointing patterns that inform the prediction of crime.

Figure 5: Burglaries in the Manchester city center in the months of April to October, 2011.



Source: Google Maps™.

6 Conclusions

Surveying concentrations of crime in time and space guides the police on how and when to prioritize the enforcement of law and order. The simplicity of the techniques used in this study enhances their usefulness. The tools employed are reusable, which allows for easy and quick analysis and display of crime data in different formats. These techniques are not limited to the area of study but are widely applicable for identifying crime patterns at a local scale, such as within cities, as well as at a global scale.

Our analysis is however limited by the exclusion of crime-influential factors such as weather and demographics. Future research should consider such factors for more precise observation and prediction. This limitation notwithstanding, the potential of crime mapping techniques to guide security-related decisions cannot be understated. Based on the results of our observation, we recommend the use of spatial mapping and visualization of crime when making predictions of future crime events.

References

[1] J. Jackson and E. Gray, "Functional fear and public insecurities about crime," *British Journal of Criminology*, 50:1-22, 2010.

[2] C. Brunsdon, J. Corcoran, and G. Higgs, "Visualising space and time in crime patterns: A comparison of methods", *Computers, Environment and Urban Systems*, 31: 52-75, 2007.

[3] M. Townsley, "Visualising space time patterns in crime: the hotspot plot", *Crime patterns and analysis*, 1:61-74, 2008.

[4] R. Frank, M. A. Andresen, and P. L. Brantingham, "Visualizing the directional bias in property crime incidents for five Canadian municipalities", *Le Géographe Canadien*, 57:31-42, 2013.

[5] L. E. Cohen and M. Felson, "Social change and crime rate trends: A routine activity approach", *American sociological Review*, 588-608, 1979.

[6] R. V. Clarke and D. B. Cornish, "Modeling offenders' decisions: A framework for research and Policy", *Crime and Justice*, 147-185, 1985.

[7] V. Ceccato and A. C. Uittenbogaard, "Space–Time Dynam-ics of Crime in Transport Nodes", *Annals of the Association of American Geographers*, 104:131-150, 2014.

[8] S. J. Rey, E. A. Mack, and J. Koschinsky, "Exploratory Space-time analysis of Burglary Patterns", Journal of Quantitative Criminology, 28:509-531, 2011.

[9] J. Ratcliffe, "Crime mapping: spatial and temporal challenges", In *Handbook of quantitative criminology*, 5-24, Springer New York, 2010.

[10] S. Chainey and J. Ratcliffe, "GIS and crime mapping", John Wiley & Sons, 2005.

[11] T. M. Davies, M. L. Hazelton, J. C. Marshall, "Package 'sparr'", *Risk*, 22:1, 2013.

[12] H. Wickham, "ggplot2: elegant graphics for data analysis". Springer New York, 2009.

[13] D, Kahle and H. Wickham, "ggmap: A package for visualization with Google Maps and OpenStreetMap". http://www.inside-r.org/packages/cran/ggmap

[14] N. Williams and N. Cowen, "Manchester riots of 2011 and the index of multiple deprivation", *Radical Statistics*, 106:30-48, 2012

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

It's GIrls' Day! What sketch maps show about girls' spatial knowledge

Vanessa Joy A. Anacta
Institute for Geoinformatics
University of Muenster
Heisenbergstrasse 2, 48149
Muenster, Germany
v.anacta@uni-muenster.de

Thomas Bartoschek
Institute for Geoinformatics
University of Muenster
Heisenbergstrasse 2, 48149
Muenster, Germany
bartoschek@uni-muenster.de

Abstract

This paper describes the analysis of sketch maps from girls who participated in the Girls' Day annual event in Germany. The event caters to girls from Grades 7 – 10 as an opportunity to experience various jobs that might interest them in the future, typically within the STEM-disciplines. One of the performed activities was asking the girls who participated to draw a sketch map of an area they are familiar with. We are interested in finding out how girls externalize the environment they were told to draw. The activity also helps us understand how they organize their environmental knowledge through sketch maps. This descriptive work deviates from gender comparison of map-making by focusing only on girls. This paper allows us to understand differences of girls' cognitive abilities based on what they have drawn on the map. The results showed that girls draw map ranging from egocentric pictorial representation with few details to survey structured map. More than 40% of the girls have included landmarks and streets outside the region of interest showing a more global view of the area. Landmarks frequently drawn showed visual, structural and cognitive characteristics. This study contributes to research related to better understanding of the cognitive abilities of young adults, particularly girls.

Keywords: spatial knowledge, sketch maps, girls, landmarks, Girls' Day

1 Introduction

People's cognitive map varies which shows different ways of how one's spatial knowledge is externalized. Some people sketch a known area with fewer details while others draw a more detailed map. Many factors could explain such differences such as environmental experience [1], spatial abilities [2,3] or gender [4]. There are various ways of how people acquire spatial knowledge. Siegel and White's [5] hypothesis show development of how people first learn the environment which is through paths. Montello [6] claim that adults already acquire landmark and configural knowledge when new in the environment. The elements which help build one's spatial knowledge had been investigated by Lynch [7]. Among them, landmarks appear to be widely used and extensively studied in the area of spatial cognition from its characteristics [8], function [9,10] and importance of location [11] specifically in wayfinding.

Development of spatial knowledge among children has long been investigated by psychologists, geographers, and cognitive scientists. Two opposing theories evolved from this research: In the constructivist approach it is being argued, that children are born without knowledge of space, and without a conception of the objects, which occupy and structure that space [12]. They construct their knowledge from the experiences they make in space. The nativist approach states that spatial understanding may be innately available to infant [13]. In the empiricism approach, spatial knowledge is primarily from sensory experience using basic minimal inbuilt capacities [14]. An adaption of these approaches could take a notion of innate abilities that are, contrary to the nativist approach, not impenetrable to each other but can be combined to create a comprehensive spatial representation of location, thus supporting the empiricist claim of using an intertwined mix of abilities to form a spatial representation that actually improves through interaction, as claimed by constructivists.

Sketch maps are spatial representations visualizing how people externalize their environment. Researchers have long analyzed cognitive aspects of sketch maps [15,16,17]. Tversky [15] investigated what sketch maps tell about how one thinks but that distortions are inevitable. On the other hand, other researchers have investigated correctness of sketch maps as it externalized what people know about the environment which could also show other important information that could not be found in metric maps [18,19]. Sketch maps have also been used in assessing what children have learned [20,21] and the differences of their cognitive abilities [22]. In this study, we focus our investigation on the girls' sketch maps based on a) mapping abilities; b) characteristics of landmarks and c) location of landmarks and streets comparing with metric maps. Results showed that more than 40% of the girls included other landmarks and streets that were not part of the region of interest. These could be helpful landmarks which they considered important to be shown for orientation purposes. This paper contributes to the study on understanding spatial knowledge of young adults.

2 Participants and Method

There were 13 girls from various schools aged 11 to 13 who participated in the annual Girl's Day event at the Institute for Geoinformatics, University of Muenster.

The girls were asked to draw any spatial feature they could remember inside the Promenade which is a bike and pedestrian lane encircling the city center. We gave them the cathedral in the center of the study area as a reference point.

They were given an A4 paper and a pen. No example was provided in order not to influence how they will draw the sketch map. They were given a maximum of 15 minutes to draw the map.

3 Results and Discussion

3.1 Mapping abilities

The girl's mapping abilities differ as shown in the sketch maps in Figure 3. Although, the task was to draw only features inside the Promenade, it showed that some of them mapped other features outside it. The inclusion of these features suggests that some people consider the importance of global features in externalizing any environment. It also showed in route maps wherein participants tend to remember and draw other landmarks both along and off the route in their sketch maps [23]. This shows that some people tend to include spatial features that will help the person orient himself/herself in the environment such as landmarks that are distant.

Figure 3 shows the different types of sketch maps some of the girls have drawn. Following Moore's [24] classification of sketch maps – Level I, Level II, and Level III, we identified similar-like classification from the girls' sketch maps. For Level I, the maps show an egocentric representation of the environment. An example is the Sketch map 1 where the participant only drew the church and some surrounding features. Sketch map 2 shows an example of Level II which is partially coordinated landmarks and streets. On the other hand, sketch maps 3 and 4 show some of the girl's survey representation of the area which could fall under Level III classification. The maps show coordinated spatial features and including other landmarks outside the study area.

3.2 Landmark characteristics

Prominent landmarks play a big role in place knowledge and navigation [8]. This is evident in most of the girls' sketch maps. The most common landmarks recalled and drawn are churches. Among all the landmarks drawn in Table 1, one is situated outside the Promenade which the girls included in their sketch map. This suggests that some girls have a global view of the environment and have considered it important to draw landmarks not only situated inside the area of interest but also those outside it which could be deemed important for orientation purposes.

Following the classification of Sorrows and Hirtle [8], the strongest landmarks in the environment showed the three properties (refer to Table 1): *visual, structural* and *cognitive*. Visual landmark refers to objects with distinct visual appearance such as the architectural design of buildings. Structural landmark pertains to the locational aspect and role of landmark in the space. Cognitive landmark, on the other hand, refers to landmark with personal meaning or importance which stands out in the environment.

Table 1: Characteristics of landmarks frequently drawn in sketch maps

Landmarks	#	Visual	Structural	Cognitive
Cathedral	11	•	•	•
Church 1	5	•	•	•
Roundabout*	5	•	•	•
City Hall	4	•	•	•
Bookstore	4	•	•	•
Church 2	3	•	•	•
Church 3	3	•	•	•
Parking Lot	3			•

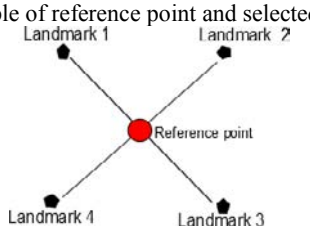
Note. The symbol # represents frequency of occurrence
**landmark outside the region of interest*

3.3 Comparison of sketch maps and metric maps

We compared with metric maps by counting the number of landmarks drawn in the sketch maps. We created a fix reference point or landmark (the cathedral) that frequently appeared in the sketch map. Figure 1 shows an example of the reference point and some selected landmarks. We selected a minimum of five landmarks in each sketch map and compared them with the correct placement in the metric map. This method is a simplified adaptation from Chipofya et al's [19] study but an extensive qualitative analysis of the landmarks and streets is beyond the scope of this paper.

In four survey type sketch maps tested, it showed that the girls have correct spatial relations of features which are close to the metric map. For instance in Figure 2, one of the girls drew landmarks outside the Promenade which showed almost correct positions when compared with the metric map. The average percentage of the four maps checked was 92.25% in terms of its correctness compared with the metric map. Sketch map 4 of Figure 3 incurred an average of 93.52% correctness where the girl drew landmarks both inside and outside the study area.

Figure 1: Sample of reference point and selected landmarks



With the cathedral as the reference point, it was easier for most girls to relate other prominent landmarks in the city as well as other distant landmarks. This relates to what Sadalla et al [25] highlighted in their study that making a prominent feature as reference point will make it easy for people to define positions of surrounding objects in space.

Figure 2: Overlaid sketch map and metric map

4 Conclusion

The sketch maps that girls drew showed differences in terms of details. Some girls have drawn less detailed sketch maps while others have drawn a survey map of the environment. In addition, some of the girls have included other spatial features that were not part of the study area. This could be for orientation purposes which they considered important to be externalized in the sketch map. This shows girls' awareness of the environment they are familiar with by including more features that will show an overall view of the area.

In comparing sketch maps with metric map, a prominent reference point played an important role in knowing the locations of adjacent spatial features in the environment which could help in the overall understanding of its spatial layout. This provides one way of knowing how to evaluate a person's knowledge of his/her environment.

This descriptive paper helps us further understand how girls visualize their environment which will develop more studies to facilitate girls' spatial thinking. For future work, spatial ability tests will be given to participants and an extensive qualitative analysis of the sketch maps will be conducted. We intend to use a recently developed drawing application for tablets which records the drawing sequence of the activity to better understand girls' spatial knowledge based on how they draw and organize the elements in the sketch map. It will also be interesting to compare sketch maps of girls across ages and cultures taking into account different experiences and exposure to maps and mapping in different educational systems.

Acknowledgement

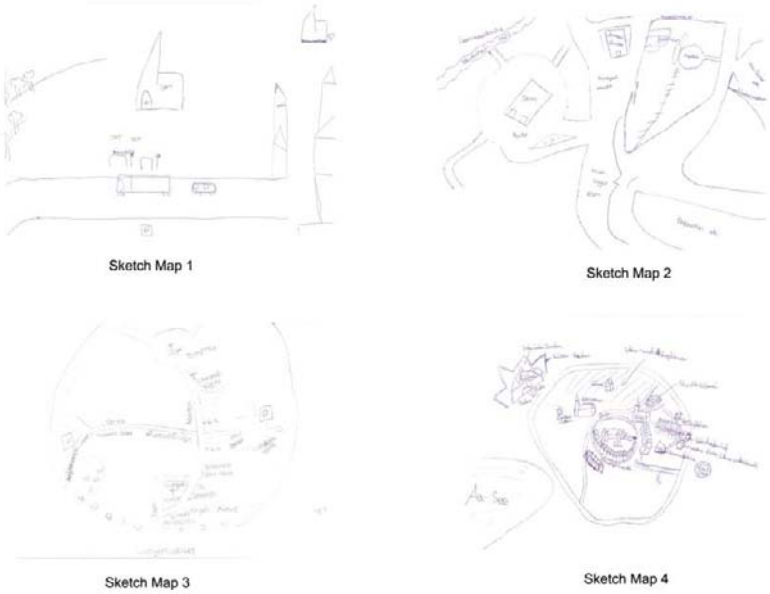
We acknowledge the support of the DAAD, GI@School and the DFG-funded WayTO with project number SCHW 1371/15-1.

References

[1] Golledge, R.: Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes. Baltimore, MD: Johns Hopkins University Press, 1999.

[2] Hegartya, M., Montello, D., Richardson, A., Ishikawa, T., Lovelace, K.: Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence* 34, Issue 2, 151-176, 2006.

Figure 3: Examples of the girls' sketch maps



[3] Montello, D., Lovelace, K., Golledge, R., Self, C.: Sex-Related Differences and Similarities in Geographic and Environmental Spatial Abilities. *Annals of the Association of American Geographers* 89, Issue 3, 515-534, 1999.

[4] O'Laughlin, E., Brubaker, B.: Use of landmarks in cognitive mapping: Gender differences in self report versus performance. *Personality and Individual Differences* 24, Issue 5, 595-601, 1998.

[5] Siegel, A., White, S.: The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior* 10, 9-55, 1975.

[6] Montello, D.: Cognitive Geography. In Thrift, R., ed. : International encyclopedia of human geography 2. Oxford Elsevier Science. (2009) 160-166

[7] Lynch, K.: Image of the City. MIT Press, Cambridge (1960)

[8] Sorrows, M., Hirtle, S.: The Nature of Landmarks for Real and Electronic Spaces. In Freksa, C., Mark, D., eds. : Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science, vol. 1661 (1999)

[9] Raubal, M., Winter, S.: Enriching Wayfinding Instructions with Local Landmarks. In Egenhofer, M., Mark, D., eds. : *GIScience '02 Proceedings of the Second International Conference on Geographic Information Science*, pp.243-259, 2002.

[10] Winter, S.: Route Adaptive Selection of Salient Features. In Kuhn, W., Worboys, M., Timpf, S., eds. : *Spatial Information Theory. Foundations of Geographic Information Science*, pp.349-361, 2003.

[11] Denis, M.: The description of routes: A cognitive approach to the production of spatial discourse. *Cahiers de psychologie cognitive* vol. 16, No.4, 409-458, 1997.

[12] Piaget, J., Inhelder, B.: The child's concenption of space. New York: Norton (1967)

[13] Spelke, E.: Nativism, empiricism, and the development of knowledge. (1998) 275-340

[14] Newcombe, N., Huttenlocher, J.: Making sense of the development of spatial cognition. *Trends in Cognitive Sciences* 5, No. 7, 316-617. Book Review, 2001.

[15] Tversky, B.: What do sketches say about thinking? In T. Stahovic, J., (Editors), R., eds. : *Proceedings of AAAI spring symposium on sketch understanding*, 2002.

[16] Taylor, H., Tversky, B.: Perspective in Spatial Descriptions. *Journal of Memory and Language* 35, 371-391,1996.

[17] Billinghurst, M., Weghorst, S.: The Use of Sketch Maps to Measure Cognitive Maps of Virtual Environments. In : *Proceedings of Virtual Reality Annual International Syposium* (VRAIS '95), 1995.

[18] Wang, J., Schwering, A.: The Accuracy of Sketched Spatial Relations: How Cognitive Errors Influence Sketch Representation. In : Presenting Spatial Information: Granularity, Relevance, and Integration. Workshop at COSIT 2009, Aber Wrac'h, France. 2009.

[19] Chipofya, M., Wang, J., Schwering, A.: Towards Cognitively Plausible Spatial Representations for Sketch Map Alignment. In : *Spatial Information Theory. 10th International Conference, COSIT 2011*, Belfast, ME, USA, September 12-16, 2011. Proceedings, vol. 6899, pp.20-39, 2011.

[20] Wise, N., Kon, J.: Assessing Geographic Knowledge with Sketch Maps. *Journal of Geography* 89, Issue 3, 123-129, 1990.

[21] Metz, H.: Sketch Maps: Helping Students Get the Big Picture. *Journal of Geography* 89, Issue 3, 114-118, 1990.

[22] Matthews, M.: Cognitive Mapping Abilities of Young Boys and Girls. *Geography* 69(4), 327-336, 1984.

[23] Anacta, V., Wang, J., Schwering, A.: Routes to Remember: Comparing Verbal Instructions and Sketch Maps. Connecting a Digital Europe Through Location and Place. *Lecture Notes in Geoinformation and Cartography*. 17th AGILE International Conference on Geographic Information Science. 2014.

[24] Moore, G.: Developmental Differences in *Environmental Cognition*. Environmental Design Research 2, 232-239, 1973.

[25] Sadalla, E., Burroughs, W., Staplin, L.: Reference Points in Spatial Cognition. *Journal of Experimental Psychology: Human Learning and Memory* 6 No. 5, 516-528, 1980.

3D Building Change Detection on the basis of Airborne Laser Scanning Data

Karolina Korzeniowska
Jagiellonian University/Institute
of Geography and Spatial Management
30-387, ul. Gronostajowa 7
Cracow, Poland
k.korzeniowska@uj.edu.pl

Norbert Pfeifer
Vienna University of Technology/Department
of Geodesy and Geoinformation
1040, Gußhaustraße 27-29
Vienna, Austria
Norbert.Pfeifer@geo.tuwien.ac.at

Abstract

The paper presents the possibility to use airborne laser scanning (ALS) for building change detection. For the analysis data was gathered during two campaigns: 2003 and 2011. As research area we chose a test site covering 0.72 km², representing different kind of land cover classes: water, buildings, vegetation, bare-ground, and other artificial and temporary objects. The extraction and classification of the objects was performed in 3D in order to preserve all information contained in the data, i.e., the original point cloud. The first, results of our study present the advantages and new possibilities in buildings change detection in 3D, which were not possible in the analysis based on satellite and aerial images only.

Keywords: LiDAR, classification, building change detection.

1 Introduction

In the last decades the detection of land cover changes (LCC) has been investigated intensively [1]. The rational is the desire to acquire the knowledge on actual anthropogenic and natural changes which occur in the environment, to understand the reason for these processes, and eventually to predict the trajectories of change in the future [2, 3].

Studies focused on the change detection of buildings and generally on LCC can be based on different kind on data. The first studies in this field were based on the analysis of old historical topographic maps, used to reconstruct changes which occurred in the past [4, 5]. With the passage of time and the development of new technologies, exploiting aerial and satellite platforms, it was possible to verify the changes on a basis of images, more precisely photographs, representing the terrain with the vertical view. Many studies are concerned with algorithms for land cover classification on basis of these images and methods for evaluation of the results [6, 7, 8].

In the last decade airborne laser scanning (ALS) matured as a technology, and it is generally used for the measurement of the Earth surface. Laser scanning provides point clouds and is inherently 3D. Often, the 3D content of the data is ignored and the point cloud is transformed during early processing stage into surfaces. In the last years, however, a plenty of different applications directly in the point cloud appeared [9]. One of these applications is LCC detection. So far this have not been analysed intensively, possibly also due to the lack of suitable time series data.

The first studies in building change detection, based on digital surface models (DSMs) comparisons, have been proposed by Murakami et al. [10], and Vögtle and Steinle [11]. Another studies [12, 13] showed the potential of integration the information from aerial images and ALS data for detection of building change. The study presented changes in 3D environment on a basis of 3D Surface Separation Map (SSM) has been proposed by Xu et al. [14].

Our aim is to investigate, if a 3D approach has advantages in building change detection for urban planning, and if such 3D approach is feasible. Due to this fact we are not focused on verification and comparison of 2D buildings classification correctness with respect to aerial and satellite images.

In this research we analysed ALS data gathered over an eight year time interval.

2 Research area and data

As research site we selected the city of Bregenz located in relatively flat terrain at the foot of the hill, and with the access to a lake. The test site covers an area of 0.7225 km². The area represent a variety of land cover classes such as: water, buildings, vegetation (forest, fields), artificial objects, temporary objects (cars, boats, trains), and bare earth terrains.

The data was gathered during two laser scanning campaigns in October 2003, and in April 2011 respectively. The data sets represent X, Y, Z coordinates and information on the number of returns for the laser shot and the return number as well as information on the intensity (reflection strength) of the point (Tab. 1).

Table 1: The data specification.

Data collection time	October 2003	April 2011
Area	850*850m	850*850m
Number of points	6911894	7433618
Density of the data	10.05	10.55
Max number of returns	2	5
Intensity	5-255	0-65534

Source: Laser scanning campaigns reports.

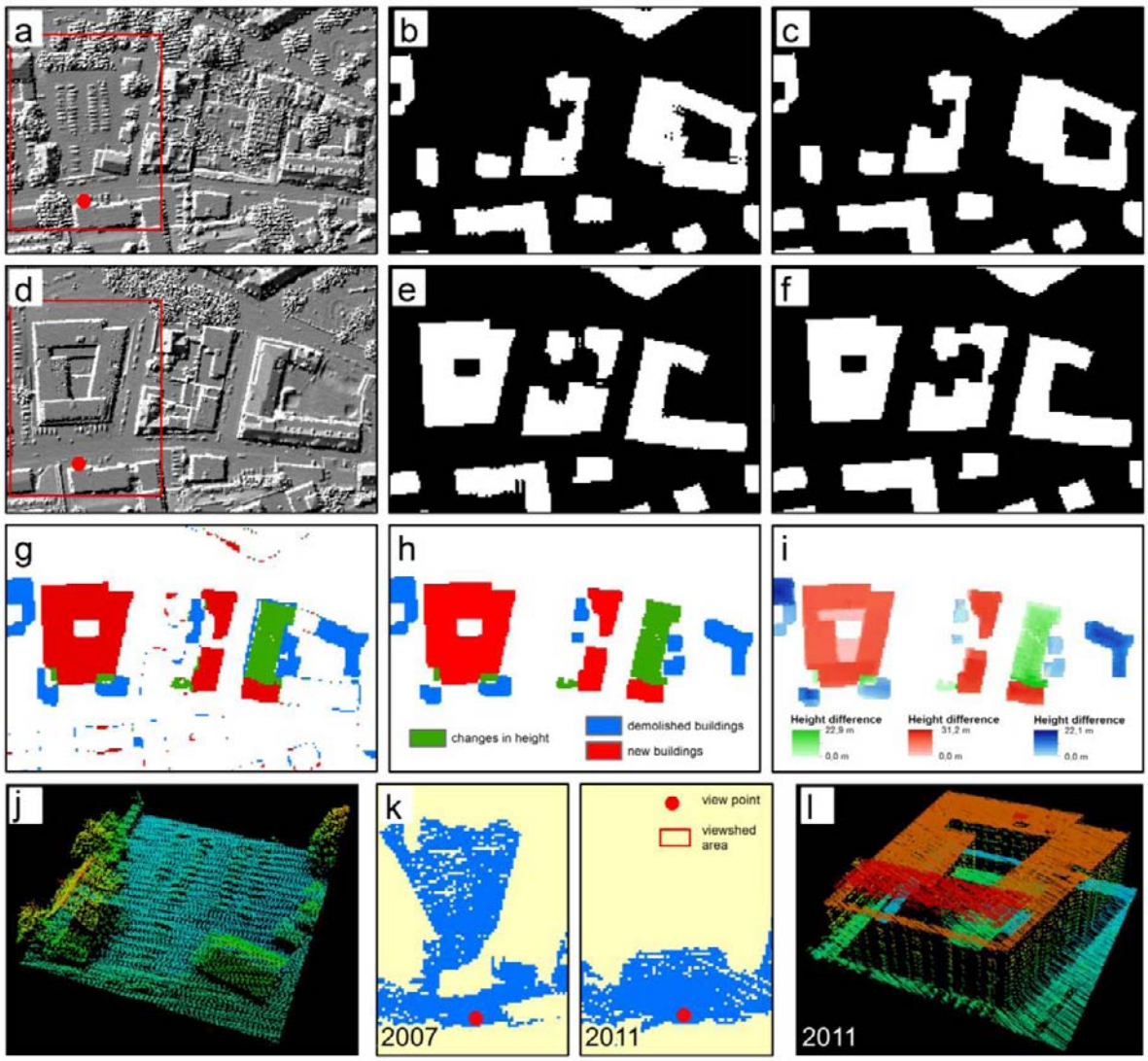
3 Applied methodology

For the data extraction and classification we applied our own automatic method based on the geometrical attributes of the data and the neighboring statistics of the extracted objects. The methodology is applied in the decision tree, on which the parameters and thresholds were selected on experience and are related to the properties of the objects. Features used were: planarity, echo ratio, normalized echo, normalized Z, and number of echoes. This was done using OPALS [15]. To preserve the information which represents the point cloud the classification has been applied in 3D. It means that as a final result we achieved the point cloud classified for ground, buildings and other data.

To assess the accuracy of the proposed method a qualitative comparison to reference data was performed. Reference data was generated by manual interpretation of the changes in the building structure for 0.31 km² area which represent similar terrain as research area in this study [16]. The quality of the classification method for the reference data is 87%, and is defined as the number of correctly classified points divided by total number of points. According to visual verification of the results we noticed that almost all the buildings were detected; however, not all the points have been correctly assigned as buildings. While the classification result is overall well a small number of classification errors were corrected manually in order to concentrate on the performance in buildings change detection.

From the classified data set we generated separately the raster digital surface model (DSM) (Fig. 1a; 1d), and digital elevation model (DEM) for both two time series data sets. From the buildings class we generated a binary map (1m pixel size) representing the building and the "not a building" classes (Fig. 1b; 1e); to remove salt and pepper noise from the images we applied morphological closing filter with kernel size equal to 1 meter (Fig. 1c; 1f).

Figure 1: Workflow



Source: Own study

The next step in our method was to extract a binary map representing building change by subtraction. The results are presented on Fig. 1g; red color represents new buildings, and blue color buildings which were demolished. To remove noise due to small errors in the alignment of the two datasets we again applied a morphological filter, this time the opening filter with 2 meter kernel size (Fig. 1.h). The kernel size in this step was bigger than in the previous step, to avoid errors which can occur, due to the X,Y coordinates shift between two data sets.

In the last step we subtract buildings height for the data sets of the two epochs and replaced them into the binary map representing new and demolished buildings. This step enabled to achieve height difference changes of the buildings (Fig. 1.i).

Additionally we evaluated changes of the view (Fig. 1.k – blue color represent visible area) that occurred in the areas marked using red rectangle in Fig. 1.a and 1.d. Smaller part of this area is also represented in Fig. 1.j – before building construction, and in Fig. 1.l - after construction of a large residential block. As a view point we selected the front of the residential block at ground level height, located in the southern part of the selected area.

4 Results

The achieved results show the usability of the ALS data for building change detection. According to our evaluation, in the area of 0.7225 km² we detect nineteen new buildings, from which eight were represented by residential buildings with a height taller than 15 meters. The analysis enables also to detect buildings which were demolished. After visual verification and comparison with the point cloud data we noticed that in our study area this phenomenon occurs for small buildings.

Due to the high accuracy of the data it was also possible to detect buildings in a renovation and buildings which shapes changes. The example is a building in Fig. 1.i. On the left side of the figure we see a changes in height in a part of new building. This change is caused by the fact that in this area in 2003 were two small, and low buildings. This buildings were demolished and on their place was built a new higher and bigger building. Different situation (not visible in the figure) we detected in eastern part of our test area where a new floor to the building was built.

The viewshed evaluated for one example building shows how, after the residential block construction, the view from selected point changed. As we see in Fig 1.k in 2011 year the northern part of the urban space is not visible.

5 Discussion and conclusions

The analysis shows the potential of the ALS data for buildings change detection. The height information contained in the data gives new possibilities in 3D change detection in time. This information can not be used on the basis of 2D satellite and aerial images. An additional dimension provides the analysis in 3D which can include expansion of build up areas in height. This information is usually used in urban planning to verify the influence of the new buildings for the sunlight reaching to the ground and also for the lighting of buildings. This is an important issue because it may result in smaller amount of light during the day inside the buildings. This can also have an effect on the temperature during sunny days, especially if newly constructed buildings on the northern hemisphere of the earth are located on the southern part of the analysed view point.

Furthermore the information about the topography relief contained in the data give us an important information regarding the topography barriers and corridors which can be determinant for further buildings changes in the future.

The results of our research provides an example of the usefulness of the ALS data for buildings change detection and new aspect – the third dimension of the data, which can be taken into consideration during the analysis.

6 References

[1] Koomen E., Stillwell J., Modelling land-use change, Theories and methods. In E. Koomen et al.(eds.). Modelling land-use change: progress and applications. Springer Verlag: 1-24, 2007

[2] Yang, Q., Li X., Shi X., 2008 Cellular automata for simulating land use changes based on supportvector machines. Computers & Geosciences, 34: 592-602.

[3] Ostapowicz K., Kozak J., 2009 Modelling of future land use change with cellular automata. In A. Car, G. Griesebner, J. Strobl (eds.) Proceedings of the Geoinformatics Forum Salzburg 2009, Wichmann Heidelberg: 158-159.

[4] Laurent G., Alain T., 2013 Three centuries of land cover changes in the largest French Atlantic wetland provide new insights for wetland conservation. Applied Geography, 42: 133-139.

[5] Cousins S. A. O., 2001 Analysis of land-cover transitions based on 17th and 18th century cadastral maps and aerial photographs. Landscape Ecology, 16(1): 41-54.

[6] Mertens B., Lambin E. F., 2000 Land-Cover-Change Trajectories in Southern Cameroon. Annals of the Association of American Geographers, 90 (3): 467-494.

[7] Nori W., Elsidding E. N., Niemeyer I., 2008 Detection of land cover changes using multi-temporal satellite imagery. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVII (B7): 947-952

[8] Tokarczyk P., Montoya J., Schindler K., 2012 An evaluation of feature learning methods for high resolution image classification. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 22nd ISPRS Congress, Melbourne, Australia.

[9] Otepka J., Ghuffar S., Waldhauser C., Hochreiter R., Pfeifer N., 2013 Georeferenced Point Clouds: A Survey of Features and Point Clouds Management. ISPRS International Journal of Geo-Information, 2: 1038-1065.

[10] Murakami H., Nakagawa K., Hasegawa H., Shibata T., Iwanami E., 1999 Change detection of buildings using an airborne laser scanner. ISPRS Journal of Photogrammetry and Remote Sensing, 54: 148-152.

[11] Vögtle T., Steinle E., 2004 Detection and recognition of changes in building geometry derived from multitemporal Laserscanning data. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 20th ISPRS Congress, Istanbul, Turkey.

[12] Rottensteiner F., 2008 Automated updating of buildings data from digital surface models and multi-spectral images: Potential and limitations. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 21st ISPRS Congress, Beijing, China.

[13] Matikainen L., Hyyppä J., Ahokas E., Markelin L., Kaartinen H., 2010 Automatic Detection of Buildings and Changes in Buildings for Updating of Maps. Remote Sensing, 2: 1217-1248.

[14] Xu S., Vosselman G., Oude Elberink S., 2013 Detection and classification of changes in buildings from airborne laser scanning data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, ISPRS Workshop Laser Scanning, Antalya, Turkey.

[15] Pfeifer N., Mandlburger G., Otepka J., Karel W., 2014 OPALS – A framework for Airborne Laser Scanning data Analysis. Computers, Environment and Urban Systems, 45: 125-136.

[16] Waldhauser C., Hochreiter R., Otepka J., Pfeifer N., Ghuffar S., Korzeniowska K., Wagner G., 2014 Automated classification of airborne laser scanning point clouds, In Solving Computationally Extensive Engineering Problems: Methods and Applications, edited by Koziel S., Leifsson L., Yang X-S., Springer, (accepted).

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Exploring twitter georeferenced data related to flood events: an initial approach

Maria Antonia Brovelli
University/Institute
Via Natta 12/14.
Como, Italy
maria.brovelli@polimi.it

Giorgio Zamboni
Politecnico di Milano
Via Valleggio 11
Como, Italy
giorgio.zamboni@polimi.it

Carolina Arias Muñoz
Politecnico di Milano
Via Valleggio 11
Como, Italy
carolina.arias@polimi.it

Alexander Bonetti
Politecnico di Milano
Via Valleggio 11 Como,
Italy
alexander.bonetti@mail.
polimi.it

Abstract

The purpose of this research-in-progress paper is to present an initial data exploration of twitter georeferenced data related to flood events, in order to determine the potential of these type of data in flood damage assessment. Data exploration aimed at determine: (1) if the features of information associated to a flood event can be generated using Twitter, (2) What are the most frequently used hash tags related to flood events? and (3) If it is possible to detect post flood information from Twitter messages. We developed a script to gather data using the Twitter Search API. The data collected gave an idea of the main feature information, spatial distribution and better search strategies.

Keywords: twitter data , flood event, damage assessment

1 Introduction

Twitter data have been playing also great role on disaster management in the last years, where research has focused mostly on disaster respond than disaster relief or post event[1]. Twitter research is quickly evolving to include more in-depth studies of social interactions and message content, that involves mostly Twitter data exploration during emergencies [2] as well as tools and methods for capturing Twitter data during and after disasters events [3]. The studies have shown that Twitter data can clearly be useful in coordinating resources and efforts and also in preparing and planning for disaster relief.

This paper presents a first step at determining the potential of Twitter georeferenced data for flood damage assessment, focusing on the following initial analysis questions: (a) What are the features of information that can be generated using Twitter associated to a flood event? (b) What are the most frequently used hash tags related to flood events? and (c) Is it possible to detect post flood information from Twitter messages?. To solve those questions, we decided to gather twitter messages with the scope of finding paths for following data analysis.

2 Twitter data collection and exploration

For twitter data collection we developed a script using PHP as a language and the Twitter Search API[1] (GET search/tweets). The script is instructed to search for a specific case-insensitive keyword, to execute queries to Twitter and to save the results in a PostgreSQL database.

The term alluvion was used for pulling Italian tweets in a half/full width forms (i.e. alluvione, alluvionato); the terms flood, inundacion, inundation and hochwasser representing the main official languages of the European Union, were also use in the query to obtain Italian tweets also in this languages. We collected Twitter data from all over the world for 69 days, from a total of 8242 unique tweet users. The results can be seen on table 1: the percentage of georeferenced tweets is very low – no more than 3% -. Among the georeferenced tweets, the ones that include an hyperlink can give very useful information, but they represent an even lower amount of the total tweets collected. The features of information contained in the hyperlinks were: online news articles, photos and videos of flooded areas, or other type of information such as maps or link to maps applications.

Table 1. Overall results

				Georeferenced Messages					
Keyword	# Messages	# GeoMessages	% GeoMessages	http Messages	Video	Photo	News	Others	No relevant info
flood	336462	7903	2,35	2192	10	643	1058	0	481
alluvion	9964	121	1,21	43	2	8	22	0	11
inundacion	6632	104	1,57	49	0	8	7	28	6
inondation	3089	91	2,95	39	0	21	14	0	4
hochwasser	565	8	1,42	4	0	1	1	0	1

Source: the authors

122 tweet messages published on Italian territory were collected, 18 of them corresponding to the word *flood* and 3 corresponding to the word *inundación*. Distribution of tweets varied over time by both distance and geographical locations thought Italy, showing a non-strong correlation between tweets and places were a flood event have occurred. The most recent big floods events near the experiment time window (26/11/2013 till 03/02/2014): were on Province of Modena (15th of Jannuary 2014) and in Sardegna island (18th of November 2013) Sardegna, but the tweets clustered on cities such as Cagliari (5 tweets), Roma (8 tweets), Milano (4 tweets). The most frequently used hash tags used by italian

[1] https://dev.twitter.com/docs/api/1.1/get/search/tweets

users were: #siamobloccati (we are blocked), #aiuto (help), #disastro (disaster), #alluvione (flood), #rischio (risk), #Sardegna, #SOS, #river.

Post flood information from Twitter messages was found mostly in the forms of news and photos that showed damage to buildings or infrastructure. No water heights information was found.

3 Conclusions and paths for future work

The data collected until now gave us an idea of the main feature information, spatial distribution and better search strategies, but further analysis are needed to determine if it is possible to detect post flood information to give a better flood event scenario.The next logical step seems to execute other queries using the most common hash tags found, but in this case using other Twitter API resources like *Streaming* or *Trends* that retrieve non only relevant tweets, but all free available tweets. Regarding data analysis can be useful to understand if there are interactions between users to find citizens networks; or to look at the message content to identify emergent themes and categories using the available data analysis software.

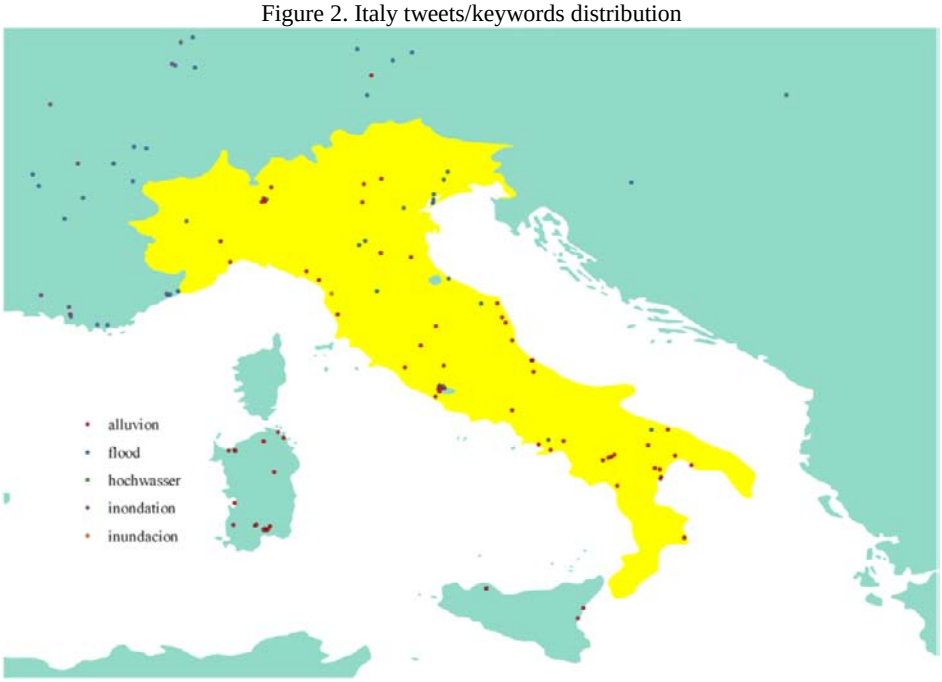
4 References

[1] M. Goodchild. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*3.3: 231-241, 2010.

[2] L. Palen. Twitter based information distribution during the 2009 Red River Valley flood threat. Bulletin of the American Society for Information Science and Technology, 2010, 36.5: 13-17.

[3] A. Bruns, # qldfloods and@ QPSMedia: Crisis communication on Twitter in the 2011 south east Queensland floods. 2012.

Figure 2. Italy tweets/keywords distribution



Source: The authors

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Texas PanHandle Climate Change Interactive GIS Web Application

Naga Raghuveer Modala
Texas A&M University
College Station
Texas, U.S.A
raghuravi23@tamu.edu

Abstract

A geographic information system (GIS) web application was developed to show the temporal and spatial variability of bias corrected historic and future climate change across the Texas PanHandle region. Spatially downscaled (50 km^2) global climate model (GCM) simulated precipitation, maximum temperature, and minimum temperature for the years 1971-2000 and 2041-2070 have been downloaded from North American Regional Climate Change Assessment Program (NARCCAP). Climate variables simulated by CRCM-CCSM, RCM3-GFDL, RCM3-CGCM3 regional climate models (RCMs) have been used for this study. Most of these model predictions are incorporated with bias due to scaling issues and immature/incomplete concepts. Typical biases include over estimation of rainfall events with low intensities and incorrect estimation of extreme temperatures. Removal of bias is important for reasonable predictions of future climate data. The bias from the historic and future climate datasets have been removed using distribution mapping technique. Bias in temperature and precipitation was removed using gaussian and gamma distribution mapping techniques respectively. This web application can be easily accessed and used by farmers, water managers, agri based industries, policy and decision makers, and other researchers to study the climate change trends across the region and plan accordingly. The web application provides the mean annual and mean monthly values of historic and future precipitation, maximum and minimum temperatures for each of the sixty seven counties in the region.

Keywords: climate change, NARRCAP, GIS, bias correction, distribution mapping technique, Texas PanHandle.

1 Introduction

The overall objective of my research was to study the impacts of future climate change on cotton lint yields in Texas PanHandle region and suggest mitigation strategies to prevent the losses due to climate change. Accessing and understanding the future climate datasets for a non-technical background people like farmers, policy decision makers from current available sources are quite challenging. The main idea of creating the interactive GIS web application was developed to provide the climate information that can be easily viewed and accessed by a common man.

Changing climate patterns will have a considerable effect on the agriculture sector, food security, water resources, and every other sector dependent on them thus affecting the regional economy. It is essential to know ahead how the climate might be changing in the region and plan accordingly to mitigate those losses.

Texas PanHandle region is the northern region of Texas encompassing 67 counties (Fig. 1). The region is a major producer of cotton and depends on precipitation and ground water pumping for its irrigation needs. This region has one of the heavily restricted ground water pumping regulations due to increasing droughts and declining ground water levels.

Figure 1: Texas PanHandle region (blue counties).

2 Methodology

Creation of the web application was a two-step process. The first step involved bias correction of the climate datasets and second step is a creation of the web application.

Three RCM simulated climate datasets were downloaded from NARCCAP. The data has a daily temporal resolution and a spatial resolution of 50 km^2. Distribution mapping technique [1] also known as statistical downscaling or quantile - quantile mapping was used to remove the bias from the climate variables. This method involves creation of a transfer function to correct the distribution function of the simulated values to match the distribution function of the observed values. This method has been successfully used in previous studies ([1]; [2]; [3])

Precipitation datasets were bias corrected using Gamma distribution mapping technique (Fig. 2) and the temperature datasets using Gaussian distribution mapping technique (Fig.

3). Before bias correcting precipitation datasets, a method has been employed to match the total number of rainfall events simulated by the RCMs with the observed data.

Figure 2: Bias correction of precipitation data using Gamma distribution mapping technique.

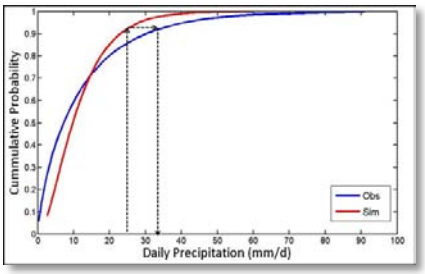
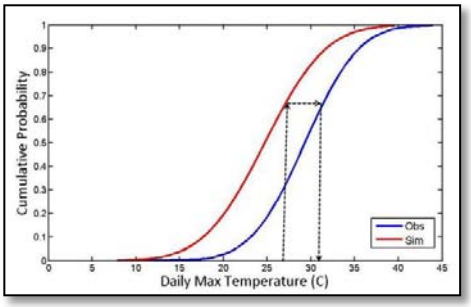


Figure 3: Bias correction of temperature data using Gaussian distribution mapping technique.



The scaling parameters that were obtained during the bias correction of model simulated historical precipitation and temperature were used to bias correct the future climate datasets.

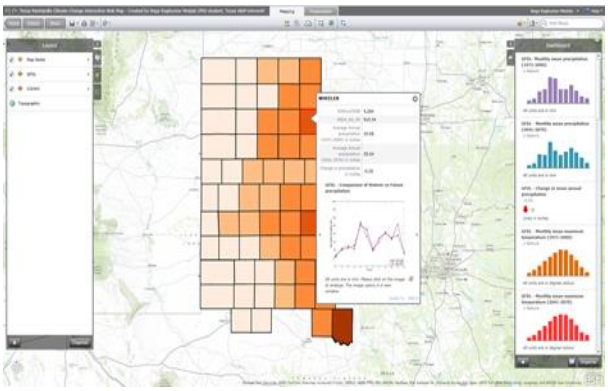
The GIS interactive web application was developed on ESRI ArcGIS Explorer platform. All the datasets for the web application were developed on ArcGIS desktop 10.1 and then uploaded them on to ArcGIS explorer online. The images generated during the bias correction process were hosted on Flickr server and then linked them to the pop-up window boxes on ArcGIS explorer online. The dashboard feature on the explorer provided a visually appealing platform to generate the statistical and summary of the climate parameters for each county. Figure 4 gives the diagrammatic representation of the methodology involved in web app development.

3 Results

Bias corrected historical precipitation and temperature data matched well with the observed data. All the models predicted an increase in temperature ranging from 2^{O}c to 4^{o}c and decrease in precipitation by 1% to 10% by 2070 for most of the counties. These changes are not uniform across the region. Only three counties showed an increase in precipitation when compared to historic data.

A web application has been developed showing the temporal and spatial variability of historic and future climate change across the region (Fig. 5). It provides a county summary climate values for the time frame of 2041-2070 at a monthly and annual scales.

Figure 5: Texas PanHandle Climate Change interactive GIS web application



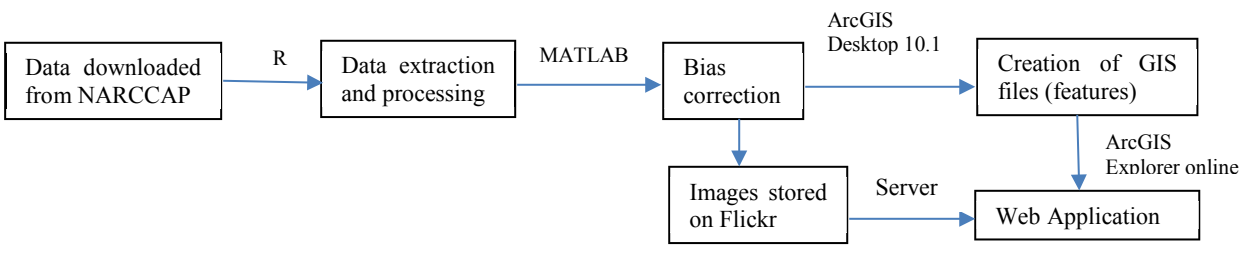
4 References

[1] Claudia Teutschbein and Jan Seibert. Bias correction of regional climate model simulations for hydrological climate-change impact studies. *Journal of Hydrology*. 456-457: 12-29, 2012.

[2] Li Haibin, Sheffield Justin and Wood, Eric F. Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *Journal of Geophysical Research: Atmospheres* (1984–2012), 115(D10), 2010.

[3] Cayan, Daniel R., Edwin P. Maurer, Michael D. Dettinger, Mary Tyree, and Katharine Hayhoe. Climate change scenarios for the California region, *Climate Change*, 87, Suppl. 1, 21–42, 2008.

Figure 4. Diagrammatic representation of methodology and tools involved in creating this web application



Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Data Scarcity or low Representativeness?: What hinders accuracy and precision of spatial interpolation of climate data?

Avit Kumar Bhowmik
Institute for Environmental Sciences, University of Koblenz-Landau
Fortstraße 7
76829 Landau in der Pfalz, Germany
bhowmik@uni-landau.de

Ana Cristina Costa
ISEGI, Universidade Nova de Lisboa
1070-312 Lisbon, Portugal
ccosta@isegi.unl.pt

Abstract

Data scarcity is a major scientific challenge for accuracy and precision of spatial interpolation of climatic fields, especially in climate-stressed developing countries. Methodologies have been suggested for coping up with data scarcity but data have rarely been checked for their representativeness of corresponding climatic fields. Here, influences of number and representativeness of climate data on accuracy and precision of their spatial interpolation were investigated and compared. Two precipitation and temperature indices were computed for a long time series in Bangladesh, which is a data scarce region. The representativeness was quantified by dispersion in the data and the accuracy and precision of spatial interpolation were computed by four commonly used error statistics derived through cross-validation. The precipitation data showed very little and sometimes null representativeness whereas the temperature data showed very high representativeness of the corresponding fields. Consequently, interpolated precipitation surfaces showed little accuracy and precision whereas temperature surfaces showed high accuracy and precision despite the scarce data. The results indicate that representativeness of climate data, i.e. variability of climate phenomenon, is more crucial than the number of data for accuracy and precision of spatial interpolation and should be treated with higher importance.
Keywords: Precipitation, temperature, point density, spatial interpolation, error statistics, regression.

1 Introduction

Spatial interpolation is an essential tool for continuously deriving climate information over space based on data at particular locations. Low accuracy and precision in spatial interpolation occurs in regions with a few climate data, e.g. in developing regions where the available number of data is often technologically and economically constrained [1, 2]. However, representativeness of corresponding climatic fields is also one of the important data characteristics and may ensure satisfactory accuracy and precision in spatial interpolation in data scare regions [3].

Consequently, the research question of this study was - can high representativeness of climate data ensure satisfactory accuracy and precision in spatial interpolation despite their scarcity?

2 Study area

In Bangladesh, only 34 meteorological stations currently report daily precipitation and temperature over 147,570 km^2 areal extent [4], and thus distinguish it as a data scarce region (Figure 1 (a)). During the period of 1948-2007, there is a gradual increase in the number of data locations for precipitation and temperature, i.e. from 8 to 32 and from 10 to 34, respectively (Figure 1(b)).

3 Materials and Methods

The daily precipitation and temperature data during 1948-2007 in Bangladesh were used. Two annual climate indices – the annual total precipitation in wet days (PRCPTOT) and the yearly maximum value of the daily maximum temperatures (TXx) were computed [5, 6].

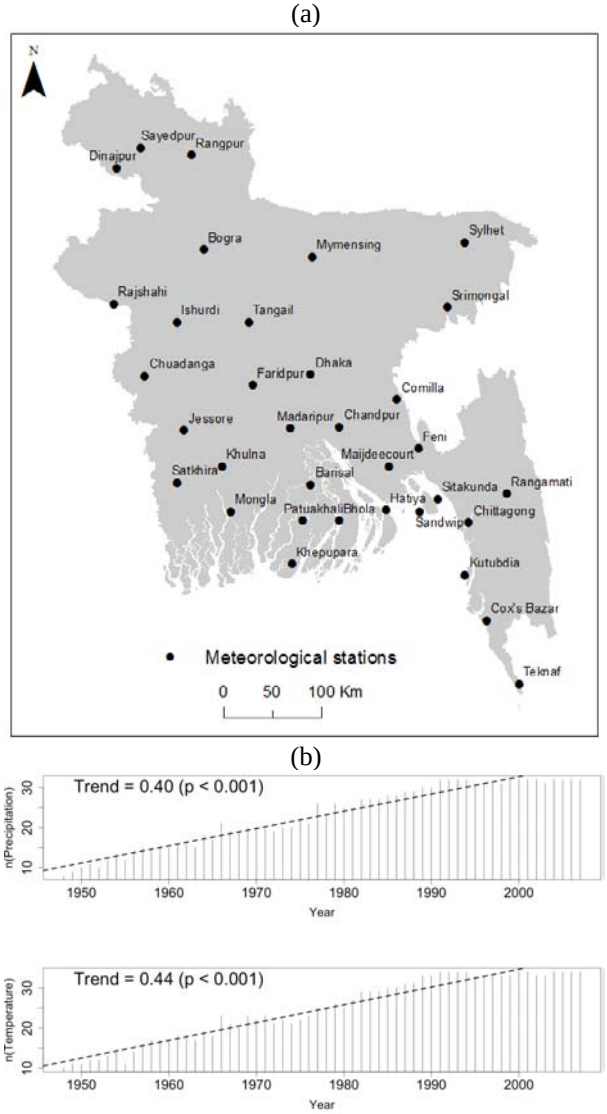
The representativeness of the climate indices was quantified by the measure of the regional coefficient of variation (k, expressed in terms of a percentage) [7].

Required number of data (n) for obtaining satisfactorily accurate and precise interpolated surfaces, i.e. root mean square error (RMSE) of the interpolated values lower than or equal to 5% of the regional mean (M) of the observed indices [8, 9], was estimated and compared to the available n according to [10, 11].

The Universal Kriging (UK) spatial interpolation model [12] was applied using the 'gstat' package [13]

in R [14]. The pooled variogram [15] parameters for three successive periods: 1948-1972, 1973-1992 and 1993-2007, taken from Bhowmik [16], were fitted to the UK model. The accuracy and precision of spatial interpolation were measured by four error statistics: (1) root mean square error (RMSE), 2) mean absolute error (MAE), 3) systematic root mean square error (RMSEs), and 4) unsystematic root mean square error (RMSEu). These were computed using the 'gstat' [13] and 'hydroGOF' [17] packages.

Figure 1: (a) Spatial distribution of the 34 current meteorological stations in Bangladesh and (b) increasing number of the data for precipitation (n(Precipitation)) (top) and temperature (n(Temperature)) (bottom) during 1948-2007.



Source: Author produced from the data [11].

The bivariate normality between the available n and the k was tested using the Henze-Zirkler's Multivariate Normality Test [18] in the 'MVN' package of R [19] and they were significantly ($p<0.001$) well fitted. The explained variability in the available n (generalized linear model with the 'poisson' family) and k (simple linear regression model) by each other were analyzed and the residuals of both models were extracted to obliterate the effects of the available n and the k on each other.

Finally, the bivariate normality in the residuals of n and k separately paired with each of the error statistics were tested [18]. Consequently, four simple linear regression models were fitted with the residuals of n and k as predictors (independent variables), separately, to predict the four spatial interpolation error statistics for both the indices (PRCPTOT and TXx). The percentage of variability in the response variables explained by the residuals of n and k was evaluated by the adjusted coefficient of determination (R^2) and through the statistical significance ($p<0.05$) of their corresponding regression parameters.

4 Results and Discussion

An average k of 41% was observed for PRCPTOT resulting from a range of 24.57-59.51% whereas for TXx the average k was 6.2% with a range of 3.26-23.97% (Table 1). Thus, the TXx data were mostly representative of the climatic field, whereas PRCPTOT data were unrepresentative.

For PRCPTOT, the number of available data did not meet the requirement for obtaining satisfactory accuracy and precision of spatial interpolation according to the computed k in any of the time steps (Table 1). On the contrary, in 43 time steps out of 60 (72%), the available n of TXx data met the requirement for satisfactory accuracy and precision.

The available n and k could significantly explain each other though the explained variability and slopes were very low (Table 2). On an average, k explains much higher variability of the error statistics than the available n and showed statistical significance when n and k were independent of each other (Table 3). The k explains 39.67% [78.16%] of the variability in the RMSE of spatial interpolation of PRCPTOT [TXx] and the corresponding regression parameters were statistically significant. Complementarily, n explains only 13.56% [3.38%] of the variability in the RMSE of PRCPTOT [TXx], thus its regression parameters are not statistically significant. More than 70% of the variability in each of the error statistics of TXx is explained by k except for RMSEu. Thus, the representativeness of the data, i.e. the variability of the

climate phenomenon, is more crucial than the number of data for ensuring accuracy and precision. The number of data is weakly related with the accuracy and precision, whereas their representativeness has a significant relation. The results are in line with [20].

Table 1: Computed representativeness (k – coefficient of variation) of the climate indices (PRCPTOT and TXx) at every time step (years) of the study period and corresponding available and required number of data (n) according to the measured k (Kelley, 2007; Lynch and Kim, 2010) for satisfactorily accurate and precise (RMSE - root mean square error <=5% of the regional mean of the indices (M)) spatial interpolation.

Years	k (%)		Available n		Required n (RMSE<=5%*M*)	
	PRCPTOT	TXx	PRCPTOT	TXx	PRCPTOT	TXx
1948	53.59*	3.69	8	10	985	10
1949	47.38	8.04	9	11	758	43
1950	54.97*	3.35	10	11	985	8
1951	53.79*	4.41	11	12	985	12
1952	37.94	6.44	10	12	423	35
1953	32.92	3.26	12	13	303	7
1954	36.71	4.38	13	14	423	11
1955	31.22	6.29	12	11	303	34
1956	24.57	6.85	14	14	137	37
1957	33.49	9.97	15	16	303	54
1958	46.61	7.52	15	17	758	41
1959	34.41	7.06	15	17	303	38
1960	42.15	5.68	15	17	573	31
1961	59.51*	4.90	16	18	985	16
1962	36.98	4.58	16	18	423	15
1963	36.05	4.67	15	17	423	15
1964	40.71	4.99	18	19	573	16
1965	42.57	7.49	17	18	573	40
1966	46.47	7.86	21	23	758	42
1967	46.17	23.97	19	20	758	137
1968	43.45	4.95	19	20	573	16
1969	44.49	8.49	20	23	573	46
1970	34.88	4.98	20	22	303	16
1971	53.34*	3.77	20	23	985	12
1972	45.54	4.98	19	21	758	16
1973	40.05	7.49	20	22	573	40
1974	43.25	5.19	20	21	573	21
1975	43.85	7.72	22	22	573	42
1976	42.05	8.38	21	23	573	45
1977	44.16	4.80	26	25	573	15
1978	39.50	4.86	23	25	423	16
1979	51.35*	5.02	26	26	985	26
1980	44.85	4.91	25	25	573	16
1981	44.74	3.67	25	27	573	10
1982	47.11	3.92	27	29	758	13
1983	42.48	4.60	27	29	573	15
1984	32.84	5.37	27	29	303	29
1985	39.56	6.76	28	30	423	30
1986	26.15	5.75	28	30	209	30
1987	36.17	5.90	29	31	423	31
1988	37.99	5.67	29	31	423	31
1989	47.31	7.24	30	33	758	32
1990	39.06	4.53	30	33	423	14
1991	37.57	4.09	32	34	423	13
1992	39.57	5.94	32	34	423	32
1993	36.75	3.97	32	34	423	13
1994	53.90*	5.64	32	34	985	30
1995	32.16	5.95	31	33	303	32
1996	38.33	5.06	31	33	423	27
1997	32.61	4.07	31	33	303	13

1998	42.04	4.35	31	33	573	14
1999	34.17	4.67	32	33	303	15
2000	49.42	3.12	32	34	758	8
2001	51.07*	3.16	32	34	985	8
2002	29.77	5.39	32	33	209	29
2003	48.09	12.81	31	33	758	69
2004	28.25	10.95	32	34	209	59
2005	37.85	5.18	32	34	423	28
2006	40.27	3.53	32	34	573	8
2007	28.55	4.15	32	34	209	13

*Null representativeness, i.e. k>50%

Source: Author produced.

Table 2: Coefficients of the simple linear regression model and generalized linear models with poisson family fitted to the representativeness (k – coefficient of variation) and the available number of data (n) of the climate indices (PRCPTOT and TXx), respectively, where were the n and k were respectively the predictors. The intercept, slope and the adjusted explained variability of the models are presented with the coefficients' statistical significance ($p<0.05$).

Response variables	Predictor variables							
	n							
	PRCPTOT				TXx			
	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
k (Simple linear regression model)	45.38*	-0.18*	0.13	1.91	7.54*	-0.05*	0.05	0.31
	k							
	PRCPTOT				TXx			
	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
n (Generalized linear model with the poisson family)	3.47*	-0008*	0.004	4.44	3.30*	-0.02*	0.009	1.98

*Statistically significant at $p<0.05$

Source: Author produced.

Table 2: Coefficients of the simple linear regression models fitted to the error statistics (RMSE – root mean square error, MAE – mean absolute error, RMSEs – systematic root mean square error and RMSEu – unsystematic root mean square error), where the representativeness (k –coefficient of variation) and the available number of data (n) of the climate indices (PRCPTOT and TXx) were separately the predictors. The intercept, slope and the adjusted explained variability of the linear regression models are presented with the coefficients' statistical significance ($p<0.05$).

Response variables	Predictor variables							
	Residuals of k							
	PRCPTOT				TXx			
Error Statistics	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
MAE	102.36*	7.91*	1.64	27.53	-0.09*	0.23*	0.02	76.29
RMSE	86.42*	11.59*	1.84	39.67	-1.09*	0.49*	0.03	78.16
RMSEs	-18.85*	9.36*	3.14	11.82	-2.21*	0.56*	0.05	72.34
RMSEu	88.44*	8.24*	2.84	11.14	-0.42	0.32*	0.03	60.88
	Residuals of n							
	PRCPTOT				TXx			
Error Statistics	Intercept	Standard slope	Standard error	R^2 (%)	Intercept	Standard slope	Standard error	R^2 (%)
MAE	579.18	-6.60	1.72	17.92	2.29	-0.04	0.01	3.22
RMSE	723.43	-7.01	2.19	13.56	3.08	-0.05	0.03	3.38
RMSEs	539.07	-7.56	3.20	7.18	2.57	-0.05	0.03	2.28
RMSEu	501.66	-3.27	2.99	0.30	2.28	-0.03	0.02	2.14

*Statistically significant at $p<0.05$

Source: Author produced.

It can be argued that the number of observations is somehow affecting the accuracy and precision of spatial interpolation of the indices, despite not being significant in general, and that its influence is considerably lower than the representativeness (Table 2) [21, 22]. Hence, in regions with abundant data, satisfactory accuracy and precision could be obtained without taking their representativeness into account [23]. However, in data scarce regions, the representativeness of the climate data should be treated with high importance.

References

[1] P. Dumolard. Uncertainty from spatial sampling: A case study in the French Alps. In H. Dobesch, P. Dumolard and I. Dyras, editors, Spatial Interpolation for Climate Data. The Use of GIS in Climatology and Meteorology, pages 57—70. ISTE Ltd., London, 2007.

[2] P. D. Wagner, P. Fiener, F. Wilken, S. Kumar, K. Schneider. Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. J. Hydro., 464-465:388-400, 2012.

[3] T. Hill and P. Lewicki. Statistics: Methods and Applications: a Comprehensive Reference for Science, Industry, and Data Mining. In StatSoft. Tulsa, OK, 2006.

[4] DMICCDMP - Disaster Management Information Center of Comprehensive Disaster Management Program. Bangladesh Meteorological Department. http://www.bmd.gov.bd/index.php. Accessed 1 May 2013.

[5] P. Frich, L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, T. Peterson. Observed coherent changes in climatic extremes during the second half of the twentieth century. Clim. Res., 19:193–212, 2002.

[6] T. C. Peterson, C. Folland, G. Gruza, W. Hogg, A. Mokssit, N. Plummer. Report WCDMP-47, WMO-TD 1071. In Report on the activities of the Working Group on Climate Change Detection and Related Rapporteurs 1998–2001. World Meteorological Organization, Geneva, 2001.

[7] M. G. Vangela. Confidence intervals for a normal coefficient of variation. Am. Stat. 50(1):21-26, 1996.

[8] J. J. Carrera-Hernández and S. J. Gaskin. Spatiotemporal analysis of daily precipitation and temperature in the Basin of Mexico. J. Hydro. 336:231-249, 2007. DOI: 10.1016/j.jhydrol.2006.12.021.

[9] P. C. Kyriakidis, J. Kim and N. L. Miller. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. J. Clim. 40(11):1855-1877, 2001.

[10] K. Kelley. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. Behav. Res. Methods. 39(4):755-766, 2007.

[11] R. M. Lynch and B. Kim. Sample size, the margin of error and the coefficient of variation. InterStat 4, 2010.

[12] R. Kerry and M. A. Oliver. Determining nugget:sill ratios of standardized variograms from aerial photographs to krige sparse soil data. Prec. Agri. 9(1-2):33-56, 2008.

[13] E. Pebesma. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30:683-691, 2004.

[14] R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna. http://www.R-project.org, Accessed 1 March 2014.

[15] R. S. Bivand, E. J. Pebesma and V. Gómez-Rubio. Applied Spatial Data Analysis with R. Springer Science and Business Media, LLC: New York, 2008.

[16] A. K. Bhowmik. A comparison of Bangladesh climate surfaces from the geostatistical point of view. *ISRN. Meteorol.*, 2012:353408, 2012.

[17] M. Zambrano-Bigiarini. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series, r package version 0.3-7. http://cran.r-project.org/web/packages/hydroGOF/, 2011.

[18] N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. Commun. Stat. Theory. Methods. 19(10):3595-3617, 1990.

[19] S. Korkmaz. MVN: Multivariate Normality Tests. http://cran.r-project.org/web/packages/MVN/index.html, 2013.

[20] B. C. Hewitson and R. G. Crane. Gridded area-averaged daily precipitation via conditional interpolation. J. Clim. 18:41-57, 2005.

[21] A. Basistha, N. K. Goel, D. S. Arya and S. K. Gangwar. Spatial pattern of trends in Indian sub-divisional rainfall. Jalv. Sam., 22:47-57, 2007.

[22] M. Radziejewski and Z. W. Kundzewicz. Detectability of changes in hydrological records. Hydrol. Sci. J., 49: 39-51, 2004.

[23] U. Haberlandt. Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. J. Hydro., 332:144-157, 2007.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

Improving equity of public transportation planning. The case of Palma de Mallorca (Spain).

Maurici Ruiz Pérez
GIS & Remote Sensing Service
University of Balearic Islands
Cra. Valldemossa Km. 7,5
07122 Palma
Spain
maurici.ruiz@uib.es

Joana Maria Seguí Pons
Earth Science Department
University of Balearic Islands
joana.segui-pons@uib.es

Jaume Mateu Lladó
Earth Science Department
University of Balearic Islands
jaume.mateu.llado@gmail.com

Maria Rosa MartínezReynés
Earth Science Department
University of Balearic Islands
mrmartinezreynes@gmail.com

Abstract

Public transport planning requires consideration of equity in access of the population to the transport service. This paper presents a methodology for the analysis of public transport in the city of Palma and its evaluation in terms of spatial and social equity. First, the analysis of supply based on the activity of the bus stops has been performed. Then an AHP multicriteria weighted model over a set of socioeconomic variables has been developed to obtain a public transport demand index. Finally the analysis of equity has been made using the Gini Index and a sensitivity analysis of bus lines. The results show the distribution of equity for all the 88 districts of city. This paper presents a simple and powerful methodology exportable to transportation planning studies in other geographic areas.

Keywords: transport equity, AHP multicriteria analysis, transport planning, gis-t

1 Background

The practice of public transport planning requires consideration of equity as an essential attribute to ensure balanced use and access of the population to the service [1] [2]. Two types of equity in transportation planning are distinguished: horizontal and vertical. The horizontal is related with spatial justice. It's oriented to maintain a balanced supply to the needs of all individuals. Vertical refers to the adjustment of supply transport to the unique needs of specific population groups (social justice).

Public transport planning requires the development of a set of tasks as: demand analysis, selection of routes and stops, setting timetabling, etc. The practice of planning requires having a deep understanding of the social situation in order to provide rational solutions adapted to the different needs of the population. In this research we have used the methodology of G. Currie [1] for public transport service analysis. The method has been adapted to our study area and a set of enhancements have been proposed: multi-criteria analysis of socioeconomic information, sensitivity analysis of bus routes. There are scarce effective methods of transport optimization based on the maintenance of social equity. In general, the demand analysis is performed giving more importance to the economic profit of the service rather than considering their social sustainability.

The main objective of this work is to propose a methodology to assess the horizontal and vertical equity of public transport and to test it at Palma municipality.

The study area is the city of Palma de Mallorca (Spain). The planning and management of public transport in Palma is performed by the Empresa Municipal de Transportes (EMT) who depends of the Palma City Council. The EMT has a total of 31 routes, involving a total of 959 bus stops distributed in an area of 19,535 hectares [4]. BUS Transport system in Palma cover the needs of a population of 421,708 people (2013) and 42,457 tourist places [3] distributed in 88 neighbourhoods. Currently its use is predominant by social groups with low income, elderly, women and students.

2 Methodology

In this research we have used the methodology proposed by G. Currie [1] for public transport service analysis. The method has been adapted to our study area and a set of enhancements have been proposed: multi-criteria analysis of socioeconomic information and a sensitivity analysis of bus routes.

2.1 Supply analysis

The analysis of the supply is based on the service level of the bus stops and includes the next steps:

- Geolocation of 959 bus stops.
- Bus stop service level (BSSL). Total buses per day are obtained for each bus stop for 12 hours period on working days.

- District Service Level (DSL). First a 300 m buffer is generated from each bus stop. Then an overlay of

$$DSLx = \sum_{n=1}^{mx} \left(\frac{Area_{Buffer\ n}}{Area_{District\ x}} * BSSL_n\right) \quad (1)$$

buffer map and districts has been made. Finally a DSL index is calculated for each district using the expression (1).
(x= district ; mx = total number of bus stops buffers included in the district x; n=bus stop)

2.2 Analysis of potential demand

We consider the total population of district as the first indicator of public transport demand.

In order to obtain a social indicator of public transport need (PTN index) for each district an AHP multicriteria analysis has been developed over a set of socioeconomic variables. It includes the next tasks:

- Creation of a socioeconomic database of Palma districts who include: demographic information, population income, economic activity, etc. All the information was provided by the Municipal Palma Observatory [5]. All the variables are normalized in a range of 0 – 1.
- A group of six experts provides weights to each of the variables according with their role in transport need. The average weight value is assigned to each variable. The final value of each district will be calculated using the expression (2).

$$PTN_x = \sum_{i=1}^{n} w_i x_i \quad (2)$$

(x= district ; i= variables ; wi = weight of variable (i); xi= value of variable (i))

2.3 Equity analysis

The supply and the demand of public transport have been analyzed jointly to identify imbalances in the transport service. The Gini index has been obtained to detect the level of inequality between public transport service and population or Public Transport Need Index.

Finally a sensitivity analysis of bus routes has been developed to detect their importance in providing equity to the bus service. For this purpose we have calculated the *Gini* index by performing modification times of each bus route.

All the cartographic information of the study can be found at the web map viewer PalmaBusTransport [6].

The software used to develop the study has been ArcGIS 10.1 and Microsoft Excel 10.

3 Results

The distribution of population in Palma shows spatial pattern of people concentration at peripheral areas. The old town has gradually lost population for the benefit of the first extension areas (*Eixample*) and the tourist area (Arenal district). There is a significant increase of population at the first belt of the city (especially toward the northeast) (Figure 1).

Regarding the social demand for transport, the result shows a significant increase at the main suburbs of Palma: Pere Garau, Bons Aires, etc. These residential areas concentrate a great number of immigrants and local population with scarce resources.

The supply of bus service has a radial distribution. The areas with the highest level of service are the main roads surrounding the old town of Palma. In these areas the bus stops have more than 200 buses per 12 h.

Equity analysis shows a distribution model with significant imbalances in different neighborhoods. Neighborhoods with low levels of horizontal equity correspond to areas with a high bus service and small population (Port Zone, El Mercado, Plaça dels Patins) or with areas of high population and smaller bus service (Pere Garau, Son Gotleu).

The Gini index for public transport service and population is 0,4 and 0,34 for public transport demand (PTD). The index increase with inequality, therefore we can say that the bus service is adapted the social needs of the city (Figure 2).

The sensitivity analysis of bus lines show that routes L15 , L3 , L7 , L8 , L33 , L5 have the greater sensitivity to maintain the level of service to the population and are responsible for maintaining the horizontal equity service (Figure 3). That means that a little change in the timetables of these routes will have a great impact on the equity of public transport system.

4 Conclusion

The proposed methodology is an improvement in the integration of various techniques of geographical and socioeconomic analysis to study the equity of public transport.

The city of Palma has a good level of bus service but moderate equity imbalances have been detected. In some populous districts of the periphery could be adequate to check the current level of bus service and adapt it to their actual level of need.

5 Acknowledgments

This research was supported by the project Civitas Dyn@mo EU-funded research project FP 7 [7].

Figure 1: Population, Public Transport Need Index and Equity Maps by districts.

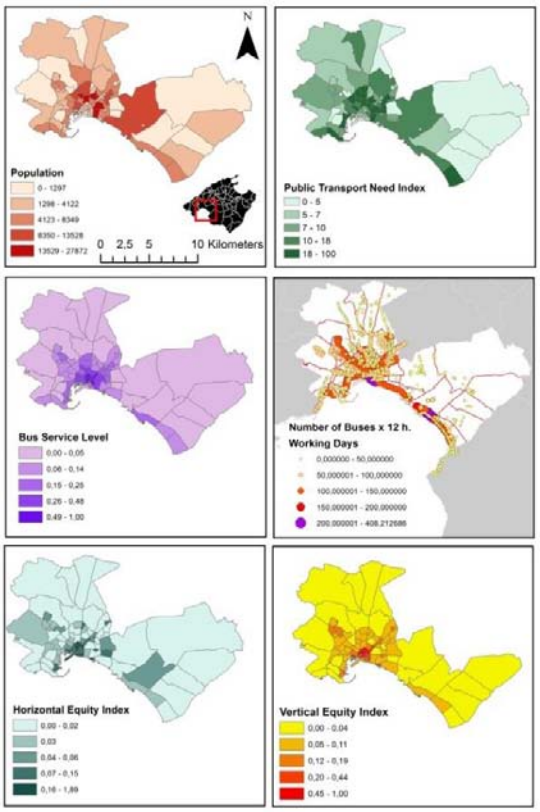


Figure 2: Lorenz Curve

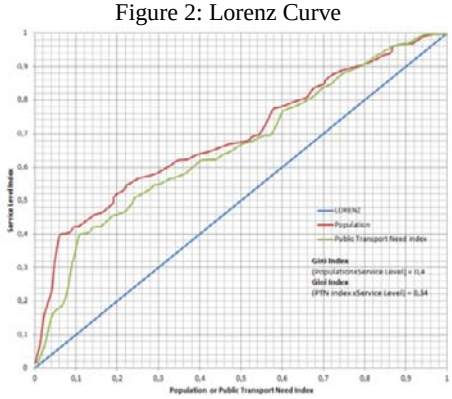
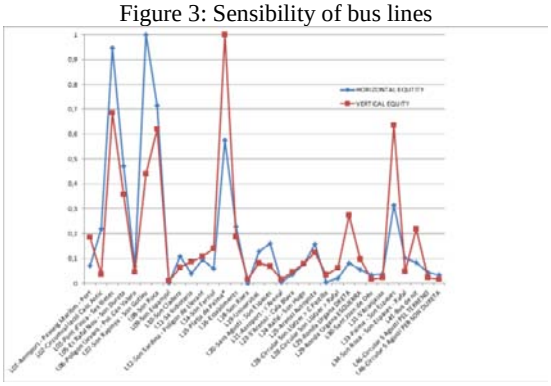


Figure 3: Sensibility of bus lines



References

[1] G. Currie. Quantifying spatial gaps in public transport supply based on social needs. *J. Transport Geography, pages*, 18(1):31-41, 2010.

[2] T. Litman. Transportation cost and benefit analysis. *Victoria Transport Policy Institute*, 1-19, 2009.

[3] http://www.ibestat.cat/ibestat/inici (last: 10/03/2014)

[4] http://www.emtpalma.es/ (last: 10/03/2014)

[5] http://observatoripalma.org (last: 10/03/2014)

[6] http://ssigt1.uib.es/flexviewers/palmabustransport (last: 10/03/2014)

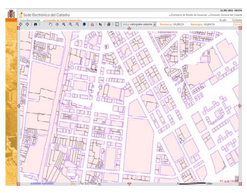
[7] http://www.civitas.eu/content/palma (last: 10/03/2014)

The CartoCiudad gamble on Open Source and value-added services

CartoCiudad is a seamless cartographic database all over Spain where all road network is topologically structured and connected. This database, built from data managed by public organizations, is the result of harmonization and integration of official digital cartography and information produced by the main stakeholders of Geographical Information in Spain.

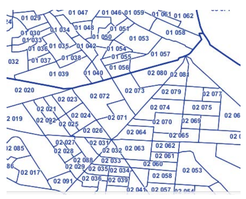
GOBIERNO DE ESPAÑA · MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS · SECRETARÍA DE ESTADO DE HACIENDA · DIRECCIÓN GENERAL DEL CATASTRO

Cadastre



- Blocks
- Building numbers
- Primitive street axis

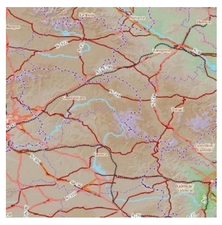
Statistical Office



- Official street names
- Wards

INSTITUTO GEOGRÁFICO NACIONAL · GOBIERNO DE ESPAÑA · MINISTERIO DE FOMENTO

National Geographic Institute



- BTN25
- Municipality boundaries
- Ortophotographies

CORREOS grupoCorreos

Post Office

- Post codes

Regional Administration

- Urban background and street axis
- Building numbers and kilometric points

CartoCiudad data are published at the geoportal www.cartociudad.es where it is possible to visualize data, locate features and calculate geoprocessing since data are handled through standard web services conform to OGC specifications. This kind of infrastructure allows users to implement cascading services and so to develop new value-added services.

Standard Web Services

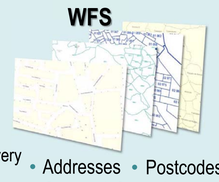
WMS / WMS-C / WMTS

- Visualize data

WPS

- Routing
- Reverse geocoding
- Influence areas definition

WFS



Discovery and download entities:
- Addresses
- Postcodes
- Census information
- Administrative units

http://www.cartociudad.es/portal/directorio-de-servicios

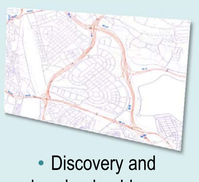
Recently a new Web Feature Service has been developed according to version 2.0.0 of the Open Geospatial Consortium (OGC) standard and complying the INSPIRE Data Specification on Addresses.

INSPIRE Web Services

WMS

- Visualize addresses and transport network

WFS

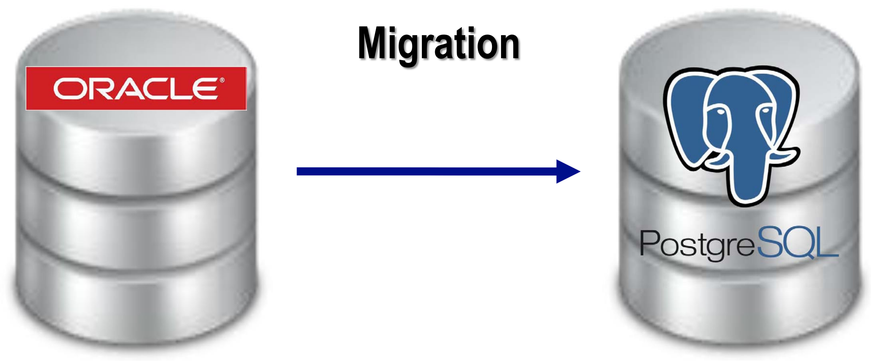


- Discovery and download addresses

http://www.cartociudad.es/wms-inspire/CARTOCIUDAD/CARTOCIUDAD

http://www.cartociudad.es/wfs-inspire/direcciones

Currently CartoCiudad database is being migrated to an open source environment: Database Management System PostgreSQL – PostGis.



Web services can be invoked by means of HTTP GET/POST requests. In addition, new functionalities demanded by end-users are available on the project site www.cartociudad.es/portal: individual or massive direct and reverse geocoding, distance calculation along the road network and WPS functionalities.

New functionalities

Direct Geocoder
addresses ⟶ X, Y

Reverse Geocoder
X, Y ⟶ addresses

Distance Calculator

WPS Client

Best Practice: AppTobaccoManagement

CMT Comisionado para el Mercado de Tabacos

A Best Practice of development of new value-added services customized according to end-user requirements is AppTobaccoManagement: application for the fulfillment of the regulation on the purchase-sale of tobacco. It bases on the location of the tobacco delivery establishments and the relative distances to the tobacco points of sale. It uses CartoCiudad services as support to manage the location, insertions and deletions of its points of interest. The main functionalities are the following:

- Point location
- Point insertion/deletion
- Candidate point of interest
- Served population

Alicia González agjimenez@fomento.es · Antonio Rodríguez afrodriguez@fomento.es · Celia Sevilla csevilla@fomento.es · Miguel Villalón mvillalon@fomento.es

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

FOODIE: Farm-Oriented Open Data in Europe

Miguel Ángel Esbrí
Atos Spain, S.A
C/Albarracín, 25, 28037
Madrid, Spain
miguel.esbri@atos.net

Abstract

The agriculture sector is a unique sector due to its strategic importance for both European citizens (consumers) and European economy (regional and global) which, ideally, should make the whole sector a network of interacting organizations. Rural areas are of particular importance with respect to the agro-food sector and should be specifically addressed within this scope. The different groups of stakeholders involved in the agricultural activities have to manage many different and heterogeneous sources of information that need to be combined in order to make economically and environmentally sound decisions, which include (among others) the definition of policies (subsidies, standardisation and regulation, national strategies for rural development, climate change), valuation of ecological performances, development of sustainable agriculture, crop recollection timing and pricing, plagues detection, etc. Such processes are very labour intensive because most parts have to be executed manually and the necessary information is not always available or easily accessible. In this context, future agriculture knowledge management systems have to support not only direct profitability of agriculture or environment protection, but also activities of individuals and groups allowing effective collaboration among groups in agri-food industry, consumers, public administrations and wider stakeholders communities, especially in rural domain.

To that end FOODIE project aims at building an open and interoperable agricultural specialized platform hub on the cloud for the management of spatial and non-spatial data relevant for farming production; for discovery of spatial and non-spatial agriculture related data from heterogeneous sources; integration of existing and valuable European open datasets related to agriculture; data publication and data linking of external agriculture data sources contributed by different public and private stakeholders allowing to provide specific and high-value applications and services for the support in the planning and decision-making processes of different stakeholders groups related to the agricultural and environmental domains.

Keywords: farm-oriented open data, VGI, Future Internet Agricultural services, sensor and remote sensing data, geospatial standards and interoperability

1 1. Introduction

The agriculture sector is a unique sector due to its strategic importance for both European citizens (consumers) and European economy (regional and global) which, ideally, should make the whole sector a network of interacting organizations. Rural areas are of particular importance with respect to the agro-food sector and should be specifically addressed within this scope.

There is an increasing tension, the like of which is not experienced in any other sector, between the requirements to assure full safety and keep costs under control, but also assure the long-term strategic interests of Europe and worldwide. In that sense, agricultural production influences, and is influenced by water quality and quantity, ecosystems, biodiversity, the economy, and energy use and supply. The seasonality and ubiquity of agriculture make agricultural practices and production amenable to efficient synoptic monitoring. Besides, food supplies depend on trends in the natural environment, including weather and climate, freshwater supplies, soil moisture and other variables. At the same time, modern agriculture has a major impact on the environment while damaging biodiversity. Unless they are sustainably managed, farms and pastures can cause erosion, desertification, chemical pollution and water shortages. These risks need to be monitored and managed by devising in effect. Therefore, from this it can be concluded that the balance between food safety and food security will be important task for future farming worldwide, but also for farming knowledge management.

The different groups of stakeholders involved in the agricultural activities have to manage many different and heterogeneous sources of information that need to be combined in order to make economically and environmentally sound decisions, which include (among others) the definition of policies (subsidies, standardisation and regulation, national strategies for rural development, climate change), valuation of ecological performances, development of sustainable agriculture, crop recollection timing and pricing, plagues detection, etc.

Such processes are very labour intensive because most parts have to be executed manually and the necessary information is not always available or easily accessible. Thus, for instance, typical farm activities carried out by farmers include the monitoring field operations, managing the finances and applying for subsidies, depending on different software applications. Farmers need to use different tools to manage monitoring and data acquisition on line in the field. They need to analyse information related to subsidies, and to communicate with tax offices, product resellers etc.

In this context, future agriculture knowledge management systems have to support not only direct profitability of agriculture or environment protection, but also activities of individuals and groups allowing effective collaboration among groups in agri-food industry, consumers, public administrations and wider stakeholders communities, especially in rural domain.

Besides, knowledge management on the agriculture domain is usually divided into three interrelated levels:

- **Macro level**, which includes management of external information (for example about market, subsidies system, weather prediction, global market and traceability systems);
- **Farm level**, which includes for example economical systems, crop rotation, decision supporting system;
- **Field (micro) level** including precision farming, collection of information about traceability and in the future also robotics.

But to exploit all these data, converted into information and finally distilled as knowledge, it is necessary to contextualize and manage this knowledge with adequate software services that assists the flow of information and synchronizes all resources and activities within a farm, making them part of farm business processes. Inventory, manufacturing, distribution, logistic, shipping, construction, and accounting processes must benefit from agriculture knowledge management to realize a new generation of ERP Software Services for modern farms, rather than using any standalone software application or any combination of them.

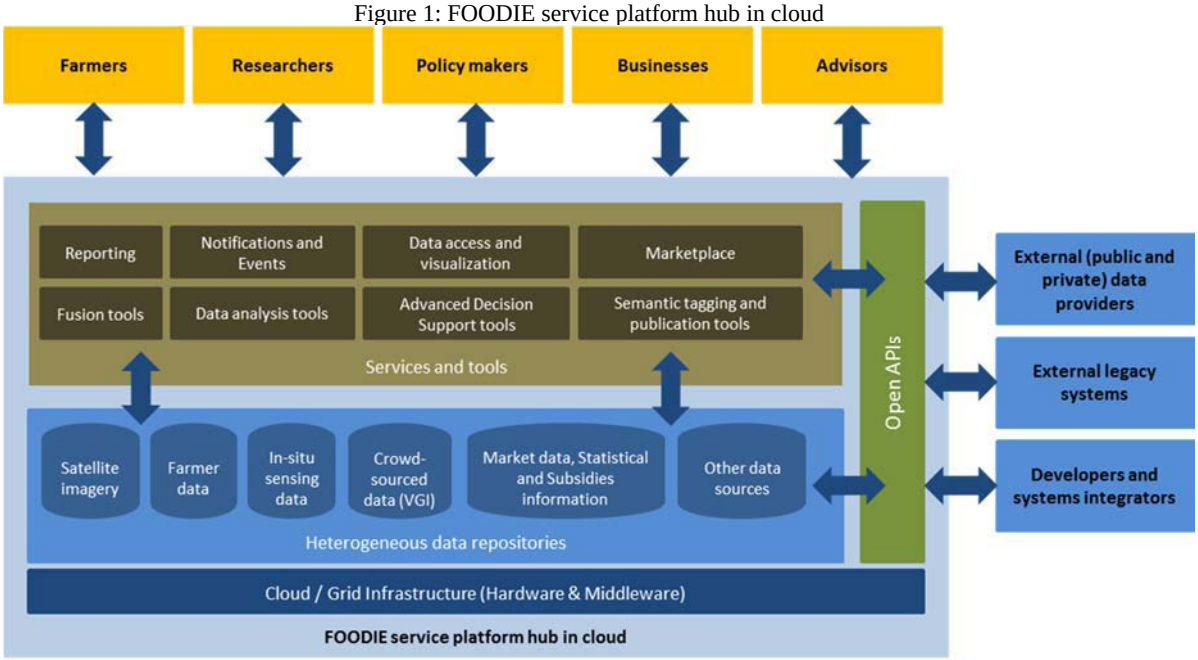
To that end FOODIE project aims at building an **open and interoperable** agricultural specialized **platform hub** on the cloud (which is conceptualized in Figure 1) for the **management of spatial and non-spatial data** relevant for farming production; for **discovery of spatial and non-spatial** agriculture related data from heterogeneous sources; **integration of existing and valuable European open datasets related to agriculture; data publication and data linking** of **external agriculture data sources contributed by different public and private stakeholders** allowing to **provide specific and high-value applications and services** for the support in the planning and decision-making processes of different stakeholders groups related to the agricultural and environmental domains.

2 FOODIE project approach

In order to realize FOODIE concept and the associated service platform hub (Figure 1), the project will aim at accomplishing the following technological objectives:

- To make use of existing spatial information resources and services for various domains –coming from different initiatives like INSPIRE,SISE, GMES/Copernicus, GNSS, GALILEO, GEOSS , GBIF, EUNIS, EEA, EUROSTAT, etc. - where the EC and the member states have invested heavily over the past decade,
- To design and provide an open and interoperable geospatial platform hub on the cloud based on existing software components from research results and available solutions in the market (mostly open-source) that includes:
  - integration of external agriculture production and food market data using principles of Open Linked Data
  - an open and flexible lightweight Application Programming Interface (API), that allows private and

Figure 1: FOODIE service platform hub in cloud



public stakeholders in the agricultural and environmental area to publish their own datasets (e.g., datasets provided by local sensor networks deployed in situ in farms, knowledge from farm communities, agricultural services companies, etc.) and make it available in the platform hub as open linked data (and enabling it to further processing and reasoning over it)

- specific and high-value applications and services for the support in the planning and decision-making processes of the different stakeholders groups
- provision of security mechanisms to prevent the unauthorised access and use of the platform users' personal information as well as the data published by them
- a marketplace where data can be discovered and exchanged but also external companies can publish their own agricultural applications based on the data, services and applications provided by FOODIE.

Besides, to facilitate integrating and deploying services over FOODIE, and trying to assure FOODIE success in the mid-term, it will be taken into account state of the art and expected evolution of management services and data marketplaces for the next years. In that sense, FOODIE will seek and provide the following innovative aspects:

- **Cloud deploying of basic and standardized services**, which will decrease not only deploying costs but also production and maintenance costs. Cloud deployment will also make easier integration and realize the vision of a "network of data-hubs", sharing data and services to provide a new data exploitation ecosystem where data is enriched by composition. Collaboration among hubs will enable a market for *data brokerage*, kind of data hub which do not store data but locates, summarizes, enrich and disaggregate data to provide vertical services of high added value.
- **Easily discoverability and composability of services**. Not only data and services published and deployed by FOODIE will follow (de facto) standards as far as possible, but guides to build and deploy services over FOODIE will be publicly available so any service can not only be easily found by end users or third party companies but also can, with the adequate access management, be reused alone or by composition with other services to provide a richer or a particular solution. This approach will also enable a personalization market realized by third-party, specialized companies in vertical markets.
- **"Pay as you go" paradigm**. Services or data published by FOODIE can be free or non-free. For instance, FOODIE will provide for free a global agriculture sector balanced scorecard and a non-free repository where key indicators for the agriculture sector may be obtained and combined by all stakeholders to make their own balanced scorecard. FOODIE may also go a step further providing analysis based on free indicators to provide free, white papers or sample reports and non-free, only for subscription members, reports. This paradigm will enable third parties as for instance consultancy companies to sell consultancy services (reports, etc.) on top of FOODIE information.
- **Reward mechanisms for data sharing**. Open data are the key value of FOODIE, but also volunteered data and knowledge shared among user's communities. FOODIE will promote participation of stakeholders and end users (high value data owners) in terms of "the more information you provide to the hub, the more data and services for free you will enjoy". Also, this approach will help to build virtual communities and exploit social knowledge.
- **Clear Return of Investment (ROI) for the end user**. The current economic situation makes reduction of costs a strategic pillar of a large number of companies. FOODIE must develop a business model which, during the marketing process, clearly demonstrate the value of services in ROI terms (i.e. FOODIE may include a simulator which calculates, asking a few questions about a crop, reduction of costs by rationalizing the use of fertilizers, water, etc. thus quickly amortizing the cost of the service)
- **Multi-device/multiplatform/multipurpose front-ends**. FOODIE will include mechanisms allowing users to exploit information and services by means of graphical and intuitive interfaces. Standards as HTML5 widgets for visualization will be preferred to assure compliance with mobility devices, as they provide automatic means to perform interface adaptation according to specific hardware and software capabilities.

3 Pilot scenarios

FOODIE concepts and objectives will be realized by means of the resulting service platform hub, which will be demonstrated in three different pilot scenarios across Europe (Spain, Czech Republic and Germany), providing each of them thus a set of common and specific requirements (from their stakeholders) in terms of data and services that will be fulfilled by the platform.

More concretely,

- **Pilot 1: Precision Viticulture (Spain)** will focus on the appropriate management of the inherent variability of crops, an increase in economic benefits and a reduction of environmental impact.
- **Pilot 2: Open Data for Strategic and Tactical planning (Czech Republic)** will focus on improving future management of agricultural companies (farms), introducing new tools and management methods, which will follow the cost optimization path, reduction of

environmental burden, improving the energy balance while maintaining production level.

- **Pilot 3: Technology allows integration of logistics via service providers and farm management including traceability (Germany)**. This pilot will focus on integrating the German Machine Cooperatives systems with existing farm management systems and logistic systems as well as to develop and enlarge existing org-cooperation models and business models with the different chain partners to create win-win situations for all of them with the help of IT solutions.

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 621074.

Huerta, Schade, Granell (Eds): Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3-6, 2014. ISBN: 978-90-816960-4-3

How Earth Observation, Crop Modeling and ICT tools can help rice cultivation: the ERMES project

Sven Casteleyn, Carlos Granell, Sergi Trilles, Joaquin Huerta
Institute of New Imaging Technologies, GEOTEC
Universitat Jaume I
Castelló de la Plana, Spain
{sven.castelyn, carlos.granell.canut, strilles, huerta}@uji.es

Mirco Boschetti, Lorenzo Busetto, Monica Pepe
Institute for Electromagnetic Sensing of the Environment
Consiglio Nazionale delle Ricerche
Milano, Italy
{boschetti.m, busetto.l, pepe.m}@irea.cnr.it

Dimitrios Katsantonis
Hellenic Agricultural Organization
DEMETER
Thessaloniki, Greece
dikatsa@cerealinstitute.gr

Roberto Confalonieri
Universita Degli Studi Di Milano
Milano, Italy
roberto.confalonieri@unimi.it

Francesco Holecz
SARMAP S.A.
Purasca, Switzerland
fholecz@sarmap.ch

Javier García Haro
Dpto. Termodinamica
Universitat de Valencia
Burjassot, Spain
j.garcia.haro@uv.es

Ioannis Gitas
Laboratory of Forest Management and Remote Sensing
Aristotelio Panepistimio Thessalonikis
Thessaloniki, Greece
igitas@for.auth.gr

Abstract

Due to pressure of food demand, increased price competition and demand for sustainable farming practices, it's increasingly important to optimize agricultural practices. The European FP7 project ERMES focuses specifically on rice cultivation, and aims to combine earth observation, crop modelling and ICT techniques and tools to optimize agro-practices and ultimately, support environmentally and economically sustainable farming systems. ERMES combines partners from Europe's three main rice-producing countries: Italy (51,3%), Spain (25,4%) and Greece (7%)[1]. With end-users and case studies in these three countries, the ERMES project is perfectly positioned to chart current practices and innovation potential in the European rice cultivation market. In this work, we focus on how ERMES plans to develop and exploit modern ICT tools and techniques to assist rice farmers to streamline the rice cultivation process on one hand, and local authorities to better regulate, control and oversee rice cultivation on the other hand.

Keywords: crop modelling, rice cultivation, smart app, geo-portal

1 Introduction

In times of continuous worldwide population growth and the progressing climate change, an efficient and sustainable food production process is more important than ever. In research areas such as agricultural engineering and agronomy, scientists dedicate their knowledge to meet these challenges. Crop modelling [1] is a primary tool developed to model, analyse and make predictions about crops (e.g., crop evolution, yield, nutrition requirements, risks, etc.), based on meteorological, environmental and crop-specific parameters. Over the years, these models were refined and complemented with satellite [2] and remote sensing data [3], whereby advances in analysis techniques play an important role [4, 5].

The European-funded FP7 project ERMES aims to apply and refine the aforementioned crop modelling techniques for rice cultivation in a European context, and combine them with modern, state-of-the-art technical solutions. According to the Food and Agricultural Organization (FAO), rice paddies are the world's second most important food commodity in terms of monetary value, and world's third in terms of quantity produced[1]. Although rice cultivation is mainly performed in Asia, European research in this area benefits from excellent available infrastructure, available technical resources and highly skilled human resources. In the ERMES project, recently available technologies, such as Sentinel-1 and -2 satellite images, technical expertise in crop modelling and remotely sensed imagery analysis, and modern ICT technologies (e.g., geo-spatial and mobile technologies), will be combined to realize improvements to current crop modelling research, and create a state-of-the-art solutions to assist farmers and local regulatory authorities in the rice cultivation process.

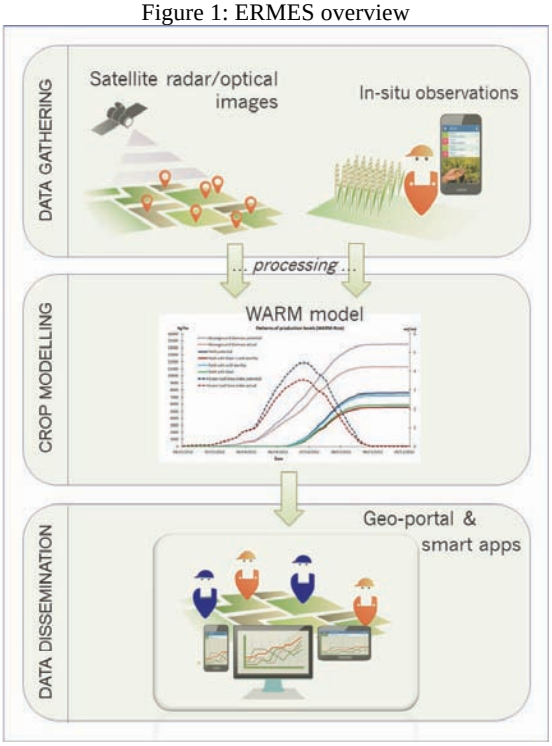
In this short paper, we give a general overview of the project, and focus on the role of modern ICT technologies within the project.

[1] http://faostat.fao.org/ ; accessed 1st of June 2014

2 ERMES in a Nutshell

Figure 1 shows a schematic overview of the ERMES project. It largely consists of three components: data gathering, crop modelling, and dissemination.

Figure 1: ERMES overview



2.1 Data gathering

The ERMES project foresees the use of three different types of data sources: 1/ traditional in-field meteorological sensor data (e.g. to measure temperature, humidity, wind speed, etc) 2/ medium and high-resolution satellite data and 3/ in-situ expert-gathered data. The advancement in our approach lies in the exploitation of modern ICT solutions to gather and analyse data, and to distribute value added information derived from the analysis to the agri-business sector. On one hand, in the European context, the availability of high resolution Sentinel-1 SAR and foreseen Sentinel-2 Optical data provides a unique opportunity to gather more precise data. On the other hand, with the proliferation of modern, cheap smart phones, packed with built-in sensors and featuring user-friendly input capabilities and continuous Internet connectivity, unique opportunities arise to gather in-situ data. In the context of the ERMES project, smart apps specifically geared towards this aim are foreseen. Finally, recent availability of low-cost multi-functional sensors may provide yet another way to further enrich data gathering.

2.2 Crop modeling

After a data processing step (i.e., quality screening, derivation of added value Earth Observation products and integration), the data resulting from the data gathering phase are used as input for a crop modeling component. The ERMES project uses the WARM rice model [6] for crop modeling and forecasting. Although this model is capable of working with incomplete input, the model accuracy and forecasting capabilities significantly improve with a more complete, and wider range of data input. By using multiple data gathering techniques, and integrating and combining the available data, we aim to provide the WARM rice model with such more complete data, and/or use complementary/overlapping data (e.g., user-provided data) to adjust and better calibrate the WARM model. Ultimately, this yields more accurate forecasting results.

2.3 Dissemination

The ERMES project targets three distinct user groups, corresponding with three distinct goals: supporting *rice farmers* for more efficient and sustainable rice cultivation, supporting *regional authorities* in controlling and overseeing rice farming practices and implementing agro-environmental policies, and providing the *agri-business* sector with relevant information.

To this aim, the ERMES project foresees two services: a regional rice service, targeted at public authorities, and a local rice service, targeted at local farmers and private companies. Based on a common Spatial Data Infrastructure (SDI), these services come in two forms:

- A Web-based geo-portal to visualize relevant geographic areas and features, overlaid with model simulations and forecasts (i.e., yield estimation, grain quality), eventual alerts related to biotic and abiotic risks (e.g., cold spells, pests), customizable information on crop development and meteorological data useful for reporting and bulletin generation, and (limited) social interactions.
- Smart apps to allow in-situ observations gathering and reporting (i.e., expert feedback), location-based information provisioning, local navigation and instant notifications (i.e., risk alerts).

The long-term goal is to transfer the solutions developed within the ERMES project, using European expertise and technical artefacts, to the global rice sector.

Acknowledgement

The ERMES project is founded in the framework of FP7-SPACE-2013 call. Contract N°: 606983. Starting Date: 01/03/2014. Duration: 36 months.

References

[1] T. R. Sinclair and N. G. Seligman. Crop Modeling: From Infancy to Maturity. *Agronomy Journal*, 88(5):698-704, 1996.

[2] G. E. Battesea, R. M. Harterb and W. A. Fullerc. An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the*

American Statistical Association, 83(401): pages 28-36. 1988.

[3] P.J. Pinter Jr., J.L. Hatfield, J.S. Schepers, E.M. Barnes, M.S. Moran, C.S.T. Daughtry, D.R. Upchurch. *Remote Sensing for Crop Modeling*. *Photogrammetric Engineering & Remote Sensing*, 69(6):647–664, 2003

[4] J.A. Richards. *Remote Sensing Digital Image Analysis*. Springer, 1999

[5] R.A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Elsevier Inc., 2007

[6] R. Confalonieri, G. Bellocchi, S. Bregaglio, M. Donatelli, M. Acutis. Comparison of sensitivity analysis techniques: A case study with the rice model WARM, *Journal of Ecological Modelling*, 221(16):1897-1906, 2011