International Journal on Advances in Systems and Measurements

2019 vol. 12 nr. 3&4

The *International Journal on Advances in Systems and Measurements* is published by IARIA.
ISSN: 1942-261x
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Systems and Measurements
Volume 12, Number 3 & 4, 2019

Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Vítor Carvalho, Minho University & IPCA, Portugal
Irinela Chilibon, National Institute of Research and Development for Optoelectronics, Romania
Soolyeon Cho, North Carolina State University, USA
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Denis Collange, Orange Labs, France
Noelia Correia, Universidade do Algarve, Portugal
Pierre-Jean Cottinet, INSA de Lyon - LGEF, France
Paulo Estevao Cruvinel, Embrapa Instrumentation Centre - São Carlos, Brazil
Marc Daumas, University of Perpignan, France
Jianguo Ding, University of Luxembourg, Luxembourg
António Dourado, University of Coimbra, Portugal
Daniela Dragomirescu, LAAS-CNRS / University of Toulouse, France
Matthew Dunlop, Virginia Tech, USA
Mohamed Eltoweissy, Pacific Northwest National Laboratory / Virginia Tech, USA
Paulo Felisberto, LARSyS, University of Algarve, Portugal
Javad Foroughi, University of Wollongong, Australia
Miguel Franklin de Castro, Federal University of Ceará, Brazil
Mounir Gaidi, Centre de Recherches et des Technologies de l'Energie (CRTEn), Tunisie
Eva Gescheidtova, Brno University of Technology, Czech Republic
Tejas R. Gandhi, Virtua Health-Marlton, USA
Teodor Ghetiu, University of York, UK
Franca Giannini, IMATI - Consiglio Nazionale delle Ricerche - Genova, Italy
Gonçalo Gomes, Nokia Siemens Networks, Portugal
Luis Gomes, Universidade Nova Lisboa, Portugal
Antonio Luis Gomes Valente, University of Trás-os-Montes and Alto Douro, Portugal
Diego Gonzalez Aguilera, University of Salamanca - Avila, Spain
Genady Grabarnik,CUNY - New York, USA
Craig Grimes, Nanjing University of Technology, PR China
Stefanos Gritzalis, University of the Aegean, Greece
Richard Gunstone, Bournemouth University, UK
Jianlin Guo, Mitsubishi Electric Research Laboratories, USA
Mohammad Hammoudeh, Manchester Metropolitan University, UK
Petr Hanáček, Brno University of Technology, Czech Republic
Go Hasegawa, Osaka University, Japan
Henning Heuer, Fraunhofer Institut Zerstörungsfreie Prüfverfahren (FhG-IZFP-D), Germany
Paloma R. Horche, Universidad Politécnica de Madrid, Spain
Vincent Huang, Ericsson Research, Sweden
Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek - Hannover, Germany
Travis Humble, Oak Ridge National Laboratory, USA
Florentin Ipate, University of Pitesti, Romania
Imad Jawhar, United Arab Emirates University, UAE
Terje Jensen, Telenor Group Industrial Development, Norway
Liudi Jiang, University of Southampton, UK
Kenneth B. Kent, University of New Brunswick, Canada
Fotis Kerasiotis, University of Patras, Greece
Andrei Khrennikov, Linnaeus University, Sweden
Alexander Klaus, Fraunhofer Institute for Experimental Software Engineering (IESE), Germany
Andrew Kusiak, The University of Iowa, USA
Vladimir Laukhin, Institució Catalana de Recerca i Estudis Avançats (ICREA) / Institut de Ciencia de Materials de Barcelona (ICMAB-CSIC), Spain
Kevin Lee, Murdoch University, Australia
Wolfgang Leister, Norsk Regnesentral (Norwegian Computing Center), Norway

Andreas Löf, University of Waikato, New Zealand
Jerzy P. Lukaszewicz, Nicholas Copernicus University - Torun, Poland
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Stefano Mariani, Politecnico di Milano, Italy
Paulo Martins Pedro, Chaminade University, USA / Unicamp, Brazil
Don McNickle, University of Canterbury, New Zealand
Mahmoud Meribout, The Petroleum Institute - Abu Dhabi, UAE
Luca Mesin, Politecnico di Torino, Italy
Marco Mevius, HTWG Konstanz, Germany
Marek Miskowicz, AGH University of Science and Technology, Poland
Jean-Henry Morin, University of Geneva, Switzerland
Fabrice Mourlin, Paris 12th University, France
Adrian Muscat, University of Malta, Malta
George Oikonomou, University of Bristol, UK
Arnaldo S. R. Oliveira, Universidade de Aveiro-DETI / Instituto de Telecomunicações, Portugal
Aida Omerovic, SINTEF ICT, Norway
Victor Ovchinnikov, Aalto University, Finland
Telhat Özdoğan, Recep Tayyip Erdogan University, Turkey
Gurkan Ozhan, Middle East Technical University, Turkey
Constantin Paleologu, University Politehnica of Bucharest, Romania
Matteo G A Paris, Universita` degli Studi di Milano,Italy
Vittorio M.N. Passaro, Politecnico di Bari, Italy
Giuseppe Patanè, CNR-IMATI, Italy
Marek Penhaker, VSB- Technical University of Ostrava, Czech Republic
Juho Perälä, Bitfactor Oy, Finland
Florian Pinel, T.J.Watson Research Center, IBM, USA
Ana-Catalina Plesa, German Aerospace Center, Germany
Miodrag Potkonjak, University of California - Los Angeles, USA
Alessandro Pozzebon, University of Siena, Italy
Vladimir Privman, Clarkson University, USA
Mohammed Rajabali Nejad, Universiteit Twente, the Netherlands
Konandur Rajanna, Indian Institute of Science, India
Nageswara Rao, Oak Ridge National Laboratory, USA
Stefan Rass, Universität Klagenfurt, Austria
Candid Reig, University of Valencia, Spain
Teresa Restivo, University of Porto, Portugal
Leon Reznik, Rochester Institute of Technology, USA
Gerasimos Rigatos, Harper-Adams University College, UK
Luis Roa Oppliger, Universidad de Concepción, Chile
Ivan Rodero, Rutgers University - Piscataway, USA
Lorenzo Rubio Arjona, Universitat Politècnica de València, Spain
Claus-Peter Rückemann, Leibniz Universität Hannover / Westfälische Wilhelms-Universität Münster / North-German Supercomputing Alliance, Germany
Subhash Saini, NASA, USA
Mikko Sallinen, University of Oulu, Finland
Christian Schanes, Vienna University of Technology, Austria
Rainer Schönbein, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Germany
Cristina Seceleanu, Mälardalen University, Sweden
Guodong Shao, National Institute of Standards and Technology (NIST), USA
Dongwan Shin, New Mexico Tech, USA
Larisa Shwartz, T.J. Watson Research Center, IBM, USA
Simone Silvestri, University of Rome "La Sapienza", Italy

Diglio A. Simoni, RTI International, USA
Radosveta Sokullu, Ege University, Turkey
Junho Song, Sunnybrook Health Science Centre - Toronto, Canada
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Arvind K. Srivastav, NanoSonix Inc., USA
Grigore Stamatescu, University Politehnica of Bucharest, Romania
Raluca-Ioana Stefan-van Staden, National Institute of Research for Electrochemistry and Condensed Matter, Romania
Pavel Šteffan, Brno University of Technology, Czech Republic
Chelakara S. Subramanian, Florida Institute of Technology, USA
Sofiene Tahar, Concordia University, Canada
Muhammad Tariq, Waseda University, Japan
Roald Taymanov, D.I.Mendeleyev Institute for Metrology, St.Petersburg, Russia
Francesco Tiezzi, IMT Institute for Advanced Studies Lucca, Italy
Wilfried Uhring, University of Strasbourg // CNRS, France
Guillaume Valadon, French Network and Information and Security Agency, France
Eloisa Vargiu, Barcelona Digital - Barcelona, Spain
Miroslav Velev, Aries Design Automation, USA
Dario Vieira, EFREI, France
Stephen White, University of Huddersfield, UK
Shengnan Wu, American Airlines, USA
Qingsong Xu, University of Macau, Macau, China
Xiaodong Xu, Beijing University of Posts & Telecommunications, China
Ravi M. Yadahalli, PES Institute of Technology and Management, India
Yanyan (Linda) Yang, University of Portsmouth, UK
Shigeru Yamashita, Ritsumeikan University, Japan
Patrick Meumeu Yomsi, INRIA Nancy-Grand Est, France
Alberto Yúfera, Centro Nacional de Microelectronica (CNM-CSIC) - Sevilla, Spain
Sergey Y. Yurish, IFSA, Spain
David Zammit-Mangion, University of Malta, Malta
Guigen Zhang, Clemson University, USA
Weiping Zhang, Shanghai Jiao Tong University, P. R. China

International Journal on Advances in Systems and Measurements
Volume 12, Numbers 3 & 4, 2019

CONTENTS

Forecasting Transportation Project Frequency using Multivariate Modeling and Lagged Variables

Alireza Shojaei
Building Construction Science Department
Mississippi State University
Mississippi State, MS, USA
e-mail: Shojaei@caad.msstate.edu

Hashem Izadi Moud
Construction Management Department
Florida Gulf Coast University
Fort Myers, FL, USA
hizadimoud@fgcu.edu

Ian Flood
M. E. Rinker, Sr. School of Construction Management
University of Florida
Gainesville. FL, USA
flood@ufl.edu

***Abstract*—Knowledge of the number of upcoming projects and their impact on the company plays a significant role in strategic planning for project-based companies. The current horizon of planning for companies working on public projects are the latest advertised projects for bidding, which in many cases is reported less than a year in advance. This provides a very short-term horizon for strategic project portfolio planning. In this research, a multivariate regression model with elastic net regularization and a support vector machine are used to forecast the Florida Department of Transportation's (FDOT) number of advertised projects in the future considering economic indices and other environmental factors. Two sets of analyses have been conducted, one with the current values of the independent variables and another one with up to 12 months lag of each independent variable. The results show that, of the predictors considered, the unemployment rate in the construction sector and the Brent oil price are the most significant variables in forecasting FDOT's future project frequency using current values. Also, it is evident that including lagged values of the independent variables increase the model's performance.**

Keywords-Multivariate Regression; Elastic Net Regularization; Strategic Planning; Project Portfolio Management; Forecasting; Support Vector Machine; Time Series.

I. INTRODUCTION

Construction companies, as with many other companies working in project-based industries, such as IT, are usually managing multiple projects concurrently while looking for new projects to maintain their business. The task of managing current (ongoing) projects while obtaining projects for continuous business is called Project Portfolio Management (PPM). A crucial part of the management of a portfolio is to make sure that the company resources and on-going projects are optimally balanced to ensure that not only each project meets its objectives but also the whole organization meets its overall goals. Management needs to make sure that they maximize the utilization of their resources by minimizing idle time while not accepting more work than they can complete effectively.

The majority of the literature focuses on internal uncertainties that pertain to PPM. In other words, the most explored aspect of the uncertainties in PPM is the relationships between the projects within the portfolio and the interaction between the current ongoing projects and possible future projects to measure their compatibility in terms of resource demand, and other criteria. However, environmental factors, such as economic conditions and specific industry conditions (for instance, oil price) can have a significant impact on a portfolio and a company's overall performance [1]. This study aims to integrate the environmental uncertainties and uncertainties regarding the unknown future projects, so that companies can apply this approach in their mid-term to long-term strategic planning. Martinsuo's [2] review of PPM frameworks showed that the uncertainty and continual changes in a company's portfolio has a significant negative correlation with its success. As a result, if users can reduce the extent of the uncertainties in their planning and have a more robust portfolio, it could greatly help their success. In summary, this paper proposes a regression model for forecasting the frequency of FDOT's future projects, which helps the user to estimate the number and timing of tendered projects in the future. The novelty of this approach is the consideration of environmental uncertainties in the model and the provision of quantitative insights into the unknown future.

The rest of this paper is organized as follows. Section II describes the impact of uncertainty on PPM and how unknown future projects can impact strategic planning. Section III describes the modeling approach followed in this paper. Section IV addresses the multivariate modeling of FDOT's number of projects in the future. Section V presents the conclusions and identifies future directions for the research.

II. Unknown Future Projects and Portfolio Strategic Planning

Planning is a vital factor in determining the success or failure of construction projects. According to Brown et al. [3] and World Bank [4] most construction projects worldwide do not meet their success targets, in terms of budget, duration or other determining factors, due to poor management practices. While success factors in different sectors of construction, like public, private, commercial, residential, infrastructure, differ, budget and equity remain as one of the main important factors that determine the success of any project. In the public sector, the federal government, as the sole client, forecasts the equity needed for the upcoming fiscal year in advance in order to accurately plan the number of needed projects to meet the society demands. Traditionally, governmental agencies had a short sighted view of the future budget; mainly due to the hardship in accurately estimating the budget needed based on the needs in the future. The process of planning future needs is costly, slow and uncertain. Also, it is usually based on the historical patterns of previously funded projects through earlier years. Using historical data for future prediction is useful, and more accurate when scope, duration, budget and type of future projects are known. Due to the unknown nature of future projects, including a lack of information about future projects' scope, number, and types, using historical data for projection purposes is not always accurate.

In principles of project management, the practice of batching multiple projects under one umbrella and defining target goals for them in a portfolio of a company is usually referred to as PPM. PPM is defined as "dealing with the coordination and control of multiple projects pursuing the same strategic goals and competing for the same resources, whereby managers prioritize among projects to achieve strategic benefit" [5]. Planview a leading Information Technology (IT) firm in project management also defines PPM as "Project portfolio management (PPM) refers to a process used by project managers and project management organizations (PMOs) to analyze the potential return on undertaking a project. By organizing and consolidating every piece of data regarding proposed and current projects, project portfolio managers provide forecasting and business analysis for companies looking to invest in new projects" [6]. PPM handles two important tasks including: (1) ensuring that the investment decisions by managing companies about the projects that participate in the portfolio are based on the single notion of maximizing the return on investment of the portfolio as a whole and minimizing the risks associated with participating projects [7], and (2) assuring that distribution of resources to different projects within the portfolio meets the portfolio goals in maximizing the portfolio and project goals and minimizing the risks [8]. Implementing an effective PPM process is challenging due to various factors involved in PPM. The golden key to a fruitful implementation of PPM in any construction enterprise is information. The future is unknown; thus having the necessary information that can paint a clear picture of the future is crucial in the PPM process. Existence of more accurate knowledge of future enables decision and policy makers to more accurately predict the future events, maximize the goals of the portfolio and projects as a whole and minimize the associated risks. This will result in maximizing the profit of the commercial enterprise [9].

The science and practice of project management is all about managing different kinds of uncertainties. Uncertainty could dramatically harm the success of any construction project regardless of the quality of staff, equipment, plans and drawings, and managers. In project management, uncertainty is defined as the degree of accuracy in determining future work processes, resource variation and work output [6][7][8]. Uncertainty is inherently coupled with risks. In traditional project management, risks and uncertainties have been usually discussed at the project level. However, it is believed that focusing on the totality of risks and uncertainties from a broader perspective might be beneficial to the success of any enterprise. While the Project Management Institute (PMI), one of the leading professional organizations in project management, discusses risks in a more general context of portfolio management, it does not provide any specific details, guidelines, plans, recommendations, directions or procedures on successfully managing future projects and portfolios uncertainties at a portfolio level. In fact, the whole concept of risk management is discussed very briefly by PMI. PMI limits discussions on different risk management to a few risk management techniques and methods and does not go beyond the management of risks and uncertainties at a portfolio level. PMI recommends only a few broad guidelines on detection, monitoring and handling uncertainties [9].

While PMI limits its discussion on risk management and uncertainties, from a scientific perspective, the best method to handle uncertainties and risks in any commercial enterprise is to analyze historical data to predict, model, project and mitigate potential harms of uncertainties and risks. At a scientific level, a variety of methods, techniques, and approaches have been tested to collect historic data, analyze the gathered data and find trends and tendencies in historical data that could help the project and portfolio managers understand the impacts of uncertainties of projects' success, and consequently portfolios. Artificial Intelligence (AI) is found to be a powerful tool in portfolio management [10]. A variety of algorithms have been developed in numerous research that can help to assess and to allocate risks and other types of uncertainties in project selection, execution and portfolio management [11]. Other contemporary analysis and computation techniques, such as multi-agent modeling [12], multi-objective binary programming [13], heuristic methods such as neural networks [14] and use of complicated Bayesian Network models [15] have been proposed and implemented by many scholars to study the nature of uncertainties, risks allocation patterns and process of risk allocation management at the project and/or portfolio levels. It is worth noting that the success rates of the aforementioned methods are not consistent. The success rates of implementing these methods vary based on numerous factors including the type of the

project, analysis method, the number of projects in a portfolio and projects and portfolio specifications. Overall, it is still mostly impossible to provide forecast models, to perfectly plan portfolios, while considering unknown projects and environmental uncertainties and risks.

III. MODELING APPROACH

The literature [5][7][14] has looked at forecasting unknown future projects with a univariate modeling approach where the number of future projects is forecasted solely based on the past values of the number of projects. This study builds upon this work by forecasting unknown future projects using multivariate regression in order to incorporate environmental uncertainties in a forecast. The data used in this case study is obtained by text mining FDOT's historical project letting database. The database covers 12 years (from 2003 to 2015) containing 2816 projects. The features extracted from the database are each project letting date, cost, and duration. Table I provides a pool of candidate independent variables including macroeconomics and construction indices compiled from the literature [5][7][14], which were available at the monthly level and did not have any missing values for the explored time frame. Table I also provides the abbreviation for each variable and the sources from which they have been obtained. These factors are considered in the regression modeling as the dependent (explanatory) variables.

The integrity and continuity of the data are important as it is a time series. As a result, random cross validation was not appropriate, and a rolling forecast origin approach was adopted for cross-validation, as illustrated in Figure 1. The data were divided into two sections, training and testing. The training period starts with three years and increases by one year in each iteration while the testing period remains steady as the three consecutive years after the training set. In other words, seven models are trained, and the average error is considered as the result of cross-validation.

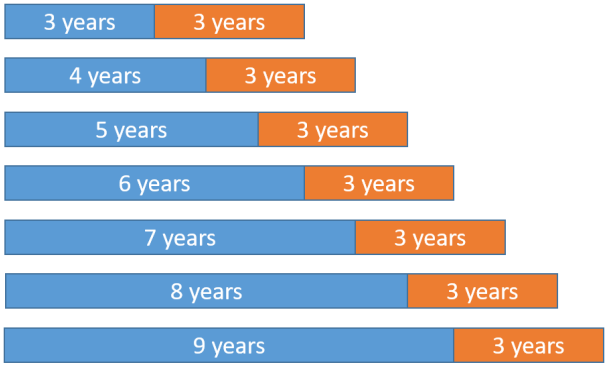


Figure 1. Forecast on a rolling origin cross-validation.

TABLE I. CANDIDATE INDEPENDENT VARIABLES.

Variable name	Abbreviation of variable	Source
Dow Jones industrial average Vol	DJI	Yahoo Finance
Dow Jones industrial average Closing	DJIC	Yahoo Finance
Money Stock M1	MS1	Federal Reserve System
Money Stock M2	MS2	Federal Reserve System
Federal Fund Rate	FFR	Federal Reserve Systems
Average Prime Rate	APR	Federal Reserve System
Producer Price Index for All Commodities	PPIACO	U.S. Bureau of Labor Statistics
Building Permit	BP	U.S. Bureau of Census
Brent Oil Price	BOP	U.S. Energy Information Administration
Consumer Price Index	CPI	U.S. Bureau of Labor Statistics
Crude Oil Price	COP	U.S. Energy Information Administration
Unemployment Rate	UR	U.S. Bureau of Labor Statistics
Florida Employment	FE	U.S. Bureau of Labor Statistics
Florida Unemployment	FU	U.S. Bureau of Labor Statistics
Florida Unemployment Rate	FUR	U.S. Bureau of Labor Statistics
Florida Number of Employees in Construction	NFEC	U.S. Bureau of Labor Statistics
Number Housing Started	HS	U.S. Bureau of Census
Unemployment Rate Construction	URC	U.S. Bureau of Labor Statistics
Number of Employees in Construction	NEC	U.S. Bureau of Labor Statistics
Number of Job Opening in Construction	JOC	U.S. Bureau of Labor Statistics
Construction Spending	CS	U.S. Census Bureau
Total Highway and Street Spending	THSS	Federal Reserve System

A. *Exploratory data analysis*

To develop the multivariate models, a better understanding of the data characteristics was first necessary, and that information was gained through an exploratory data analysis and the identification of potentially relevant predictors.

The first exploratory analysis consisted of correlation analysis. Figure 2 provides the correlation plot of the variables. The color indicates the magnitude of the correlation, and the direction of the ellipse illustrates the direction of the relationship. Furthermore, the concentration of the ellipse tells us about the degree of the linear relationship between the variables. Project frequency is represented by "freq" in the last row and column. It appears that none of the exploratory variables had a strong linear relationship with the project frequency.

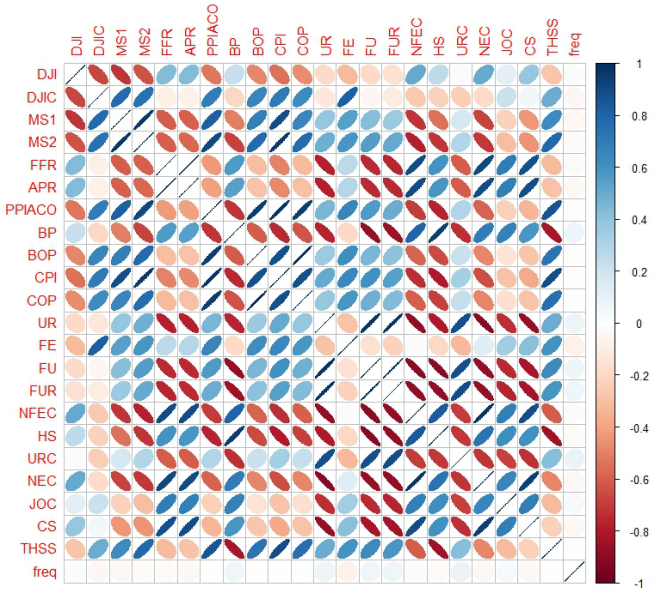


Figure 2. Correlation plot.

B. *Feature selection and feature importance*

Feature selection is the process of selecting the most relevant predictors and removing irrelevant variables from the pool of potentially useful predictors. Depending on the model's structure, feature selection can improve a model's accuracy. This process can be carried out by measuring the contribution of each variable to the model's accuracy, and then removing irrelevant and redundant variables while keeping the most useful ones. In some cases, irrelevant features can even reduce a model's accuracy. In general, there are three approaches to feature selection: the filter method, wrapper method, and embedded method.

Embedded methods implement feature selection and model tuning at the same time. In other words, these machine learning algorithms have built-in feature selection elements. Examples of embedded method implementations include LASSO and elastic net. Regularization is a process in which the user intentionally introduces bias into the training, preventing the coefficients from taking large values. This method is especially useful when the number of variables is high. In such a situation, the linear regression is not stable and in which a small change in a few variables results in a large shift in the coefficients. The LASSO approach uses L1 regularization (adding a penalty equal to the magnitude of the coefficient), while ridge regression uses L2 regularization (adding a penalty equal to the square of the magnitude of the coefficient). The elastic net uses a combination of L1 and L2. Ridge regression is effective in reducing a model's variance by minimizing the summation of the square of the residuals. The LASSO method minimizes the summation of the absolute residuals. The LASSO approach produces a sparse model that minimizes the number of coefficients with non-zero values. As a result, this approach has implicit feature selection. The generalized linear method implemented in the next section uses an elastic net. This approach incorporates both L1 and L2 regularization and thus has implicit feature selection.

Feature reduction methods, such as principal component analysis (PCA), are widely used in studies to reduce the number of independent variables. The output of such methods is a reduced set of new variables extracted from the initial variables while attempting to maintain the same information content. However, using these methods can drastically decrease the ability to interpret the significance of each input, which in itself can be very beneficial. For example, in this study knowing that oil price has a significant impact on the frequency of the projects compared to construction spending can provide valuable insight both for policy makers and contractors. As a result, the authors have chosen not to implement feature reduction methods, such as PCA.

Looking at the correlation between independent variables and the dependent variable, it became evident that a filter method using a correlation analysis was not useful, as all the variables had a nonsignificant relationship with the project frequency. As a result, an elastic net approach is used in the next section.

IV. MULTIVARIATE MODELING

The general process of model optimization and feature selection consisted of first defining a set of model parameter values to be evaluated. Then, the data was preprocessed in accordance with a 0-1 scale to make sure the high value in some variables are not skewing the model's coefficient and other variables' importance. For each parameter set, the cross-validation method discussed earlier served to train and test the model. Finally, the average performance was calculated for each parameter set to identify the optimal values for the parameters.

Ordinary linear regression is based on the underlying assumption that the model for the dependent variable has a

normal error distribution. Generalized linear models are a flexible generalization of the ordinary linear regression that allows for other error distributions. In general, they can be applied to a wider variety of problems than can the ordinary linear regression approach. Generalized linear models are defined by three components: a random component, a systematic component, and a link function. The random component recognizes the dependent variable and its corresponding probability distribution. The systematic component recognizes the independent variables and their linear combination, which is called the linear predictor. The link function identifies the connection between the random and systematic components. In other words, it pinpoints how the dependent variable is related to the linear predictor of the independent variables.

Ridge regression uses an L2 penalty to limit the size of the coefficient, while LASSO regression uses an L1 penalty to increase the interpretability of the model. The elastic net uses a mix of L1 and L2 regularization, which makes it superior to the other two methods in most cases. Using a combination of L1 and L2, the elastic net can produce a sparse model with few variables selected from the independent variables. This approach is especially useful when multiple features with high correlations with each other exist.

A generalized linear model was fit to the data at the current values using the cross-validation method discussed earlier. Alpha (mixing percentage) and lambda (regularization parameter) were the tuning parameters. Alpha controls the elastic net penalty, where $\alpha=1$ represents lasso regression, and $\alpha=0$ represents ridge regression. Lambda controls the power of the penalty. The L2 penalty shrinks the coefficients of correlated variables, whereas the L1 penalty picks one of the correlated variables and removes the rest. Figure 3 illustrates the results of the generalized linear model (for each set of parameters 7 models according to cross-validation method is trained and the average error is assigned to the set of parameters under study), optimized by minimizing the RMSE with controlling alpha and lambda. The optimized parameters were $\alpha=1$ and $\lambda= 0.56$. The authors also tested λ higher than 0.56 up to 1, however, the coefficients were not well-behaved beyond lambda=0.56.

Figure 4 depicts the LASSO coefficient curves. Each curve represents a variable. The path for each variable demonstrates its coefficient in relation to the L1 value. The coefficient paths more effectively highlight why only two variables were significant in the generalized linear model. When two variables were excluded, all other coefficients became zero at the L1 normalization, and this arrangement yielded the best performance. Figure 5 offers the variable importance for the generalized linear model with all the variables. Only the unemployment rate in the construction industry, the Brent oil price, and the unemployment rate (total) had non-zero coefficients. However, the unemployment rate (total) seemed to be relatively insignificant.

To further prune the generalized linear model, another model with only the unemployment rate in the construction sector and the Brent oil price was trained and tested. Table II contains the optimized parameters (coefficients and intercept) for the generalized linear models. The general unemployment rate had a low coefficient and, upon pruning it, the authors saw an improvement in the performance of the model. The most important variable was the unemployment rate in construction having the highest coefficient of 4.03.

Table III illustrates the performance of the optimized general linear model using a different dataset on the cross-validation sections. It was evident that excluding the unemployment rate improved the model's performance over most of the cross-validation data sections. It is notable that the pruned model performed much better in data section 1 which had the highest error and produced a more evenly distributed error among the different data sections tested. The only variables contributing to the final linear model were the unemployment rate in the construction sector and the Brent oil price.

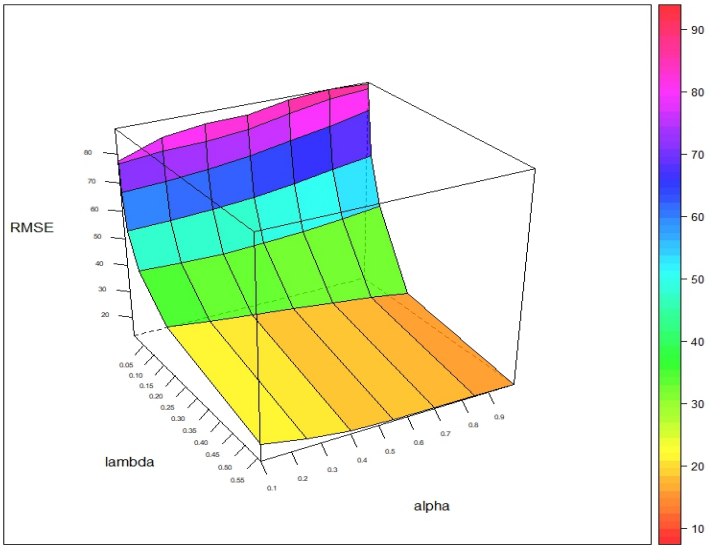


Figure 3. Generalized linear method optimization.

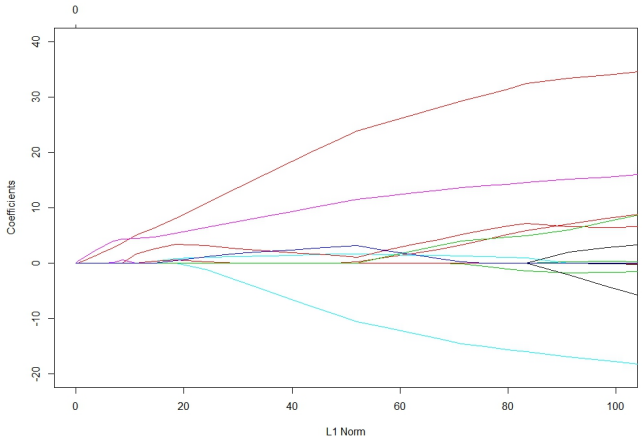


Figure 4. Lasso coefficient curve.

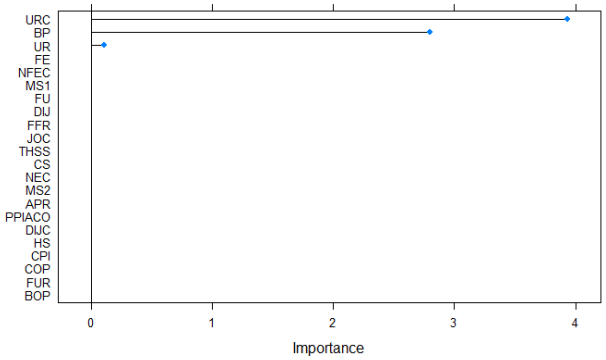


Figure 5. Variable importance of the generalized linear model.

TABLE II. PARAMETERS OF THE GENERALIZED LINEAR MODEL AT CURRENT VALUES.

Variables	Coefficients	Coefficients (Pruned by one variable)
URC	3.94	4.03
BP	2.80	2.77
UR	0.11	-----
Intercept	17.14	17.16

TABLE III. PERFORMANCE OF THE GENERALIZED LINEAR MODEL AT CURRENT VALUES.

Error term	RMSE		MAE	
Cross-validation set	**All**	**Pruned**	**All**	**Pruned**
1	16.13	9.78	13.24	10.8
2	11.58	11.94	9.64	8.56
3	13.86	13.69	11.6	8.01
4	13.16	13.14	10.82	8.25
5	12.07	10.94	9.55	10
6	11.03	10.27	8.53	8.6
7	10.89	10.87	8.6	11.28
Average	12.67	11.52	10.28	9.36

A Support Vector Machine (SVM) with a Radial Kernel were also trained and tested using the cross-validation method adopted in this study to evaluate the possible nonlinear relationship between the variables. Figure 6 depicts the results of the parameter optimization of the SVM model optimized by minimizing the RMSE with controlling sigma and C. The optimal parameters selected were sigma = 0.211 and C= 0.5.

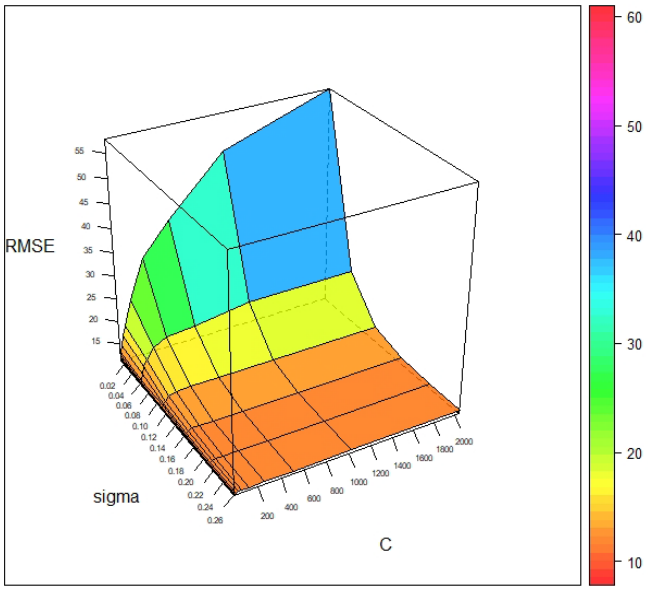


Figure 6. SVM parameter optimization.

Table IV presents the performance of the optimized SVM model on the test sets of the cross validations data sets. The results of the SVM are better than the GLM model considering all the variables at the current values.

The two GLM and SVM models so far were trained and tested on the current values of the independent variables regarding each instance of the project frequency. However, some social and economic indices might impact the dependent variable with some lag, which means that a change in the oil price might take three months to have an impact on the number of projects that FDOT is going to advertise. Figure 7 depicts the possible relationships between the variables. In the next step of this study, GLM and SVM models were trained and tested on each independent variables' current value and past 12 months values to test for both linear and nonlinear relationships between the lags of the independent variables and project frequency.

TABLE IV. PERFORMANCE OF THE SVM MODEL.

Cross-validation set	RMSE	MAE
1	10.93	8.76
2	10.31	8.25
3	9.94	7.19
4	12.06	9.63
5	11.95	9.24
6	11.11	8.38
7	10.98	8.61
Average	11.04	8.58

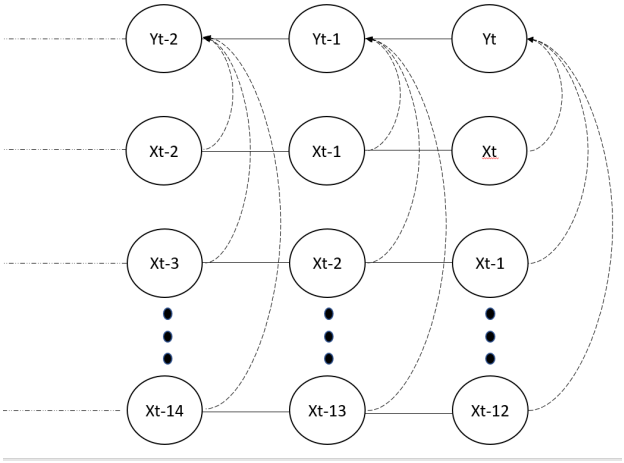


Figure 7. The possible relationship between the variables.

Figure 8 illustrates the results of the generalized linear model, optimized by minimizing the RMSE with controlling alpha and lambda over all the variables with their lagged values. The optimized parameters were α=1 and λ= 3.10. Figure 9 depicts the LASSO coefficient curves of the GLM model. Each curve represents a variable. The path for each variable demonstrates its coefficient in relation to the L1 value. The nature of the lagged value variables make them highly correlated to each other, and as a result, the L1 regularizations removes most of the variables in the process. Table V presents the results of the GLM model with the lagged variables. On the one hand, a comparison of the results with the GLM model's results including only the current values showed that including the lagged variables can increase the performance of the model. On the other hand, GLM is not friendly to variables with high correlation, and other linear models might show higher accuracy for this problem.

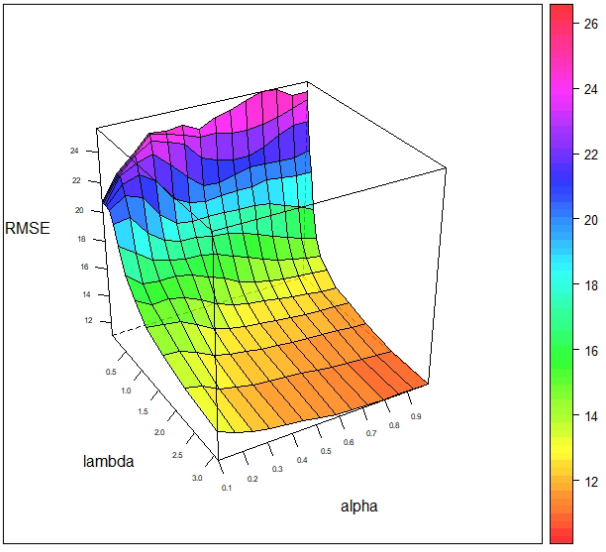


Figure 8. GLM parameter optimization with lagged variables.

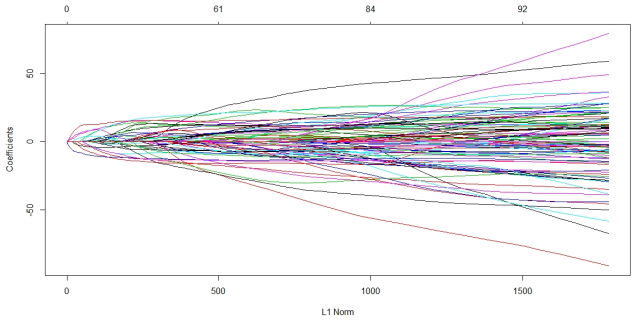


Figure 9. Lasso coefficient curves for GLM with all the variables with lagged values.

TABLE V. PERFORMANCE OF THE GLM MODEL WITH THE LAGGED VARIABLES.

Cross-validation set	RMSE	MAE
1	10.34	8.31
2	12.08	9.65
3	9.92	7.27
4	11.98	9.25
5	11.07	8.43
6	11.07	8.61
7	10.95	8.68
Average	11.06	8.60

To test the nonlinear relationship between the lagged variables and the project frequency an SVM was trained and tested using the same cross validation method. Figure 10 depicts the results of the parameter optimization of the SVM model optimized by minimizing the RMSE with controlling sigma and C. The optimal parameters selected were sigma = 0.004 and C= 0.05. Table VI presents the performance of the optimized SVM model on the test sets of the cross validations data sets. The results of the SVM are very close to the GLM with the lagged variables and SVM with only the current values. As a result, adding the lagged values increased the performance of the GLM close to the SVM model but did not increase the performance of the SVM model.

TABLE VI. PERFORMANCE OF THE SVM MODEL WITH LAGGED VARIABLES.

Cross-validation set	RMSE	MAE
1	11.06	8.22
2	10.22	7.31
3	11.90	9.36
4	11.86	9.03
5	10.99	8.71
6	10.48	8.19
7	11.19	8.61
Average	11.10	8.49

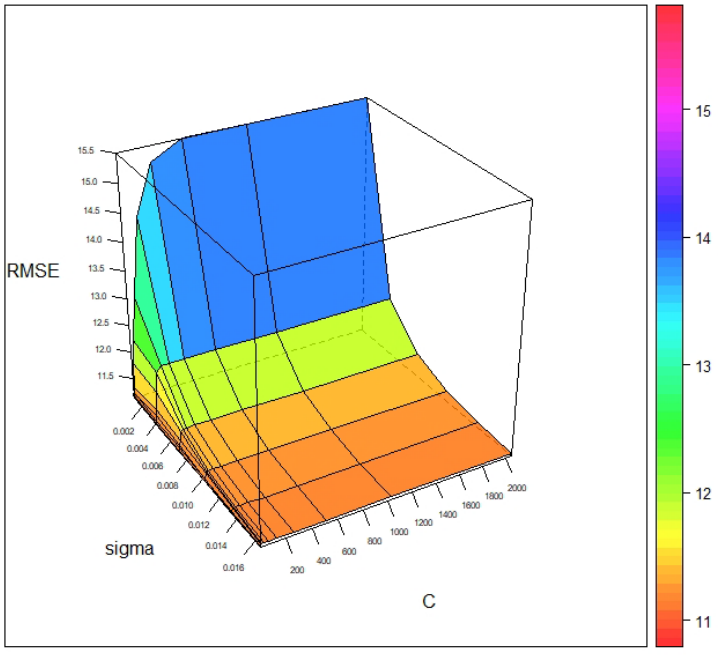


Figure 10. SVM parameter optimization with lagged variables.

Table VII provides a comparison between the multivariate models proposed in this study and some other univariate models studied previously by the authors [9]. Comparing the error terms shows that the multivariate models did not outperform some of the univariate models, such as Autoregressive Moving Average (ARMA). However, it comes close to the best performing example and it provides insight regarding the impact of environmental uncertainties on future project streams and thus could be valuable in long term strategic planning.

TABLE VII. PERFORMANCE COMPARISON OF DIFFERENT MODELS.

Model	RMSE	MAE
GLM Regression with Current variables	11.52	9.36
SVM with Current variables	11.04	8.58
ARMA(8,8)	10.71	8.45
ARMA(12,12)	11.55	9.23
AR(8)	10.92	8.48
Exponential MA (8)	11.4	9.02
GLM regression with lagged Variables	11.05	8.59
SVM with lagged Variables	11.09	8.49

It is important to note that the result of these models is the frequency of FDOT's unknown future projects, about which the user would otherwise have no information. Having reliable estimates with known error margins regarding unknown future projects can arguably provide more insight in strategic planning for a company's future compared to the current conjecture-based decision making. It should be noted that the accuracy of the models as long as the models are stable (the error is not systematic but random) is acceptable. These models are forecasting an unknown-unknown variable in the future for which there is no information available regarding their existence. However, users can use the output of this model including the error margin as inputs to their strategic planning.

The output of this research can provide a quantitative insight as a foundation for future planning. It should be noted that this model is not a standalone portfolio management framework, rather it is a supplement to existing models. For example, knowing that there is likely to be a decrease or increase in the number of projects in the future can help a company prepare in terms of consolidating or expanding its resources and assets. Furthermore, this study is limited to the FDOT's project letting database and applicability of the concept of looking into unknown-unknown projects in the future using historical data should be tested on other datasets in future work.

V. CONCLUSION AND FUTURE WORK

The importance and impact of upcoming projects on a project portfolio have been established in previously published work. However, little work has been done considering the uncertainties regarding incorporating unknown future projects in long term strategic planning. In this paper, an approach for incorporating environmental uncertainties for forecasting the number of unknown future projects is presented. Two multivariate models, generalized linear regression with elastic net regularization and support vector machine were used to forecast FDOT's unknown future projects using economic and construction indices, once with current values and once with both current and lagged values. The results indicate that the approach can reduce the impact of uncertainties on a portfolio and thus enable the development of a more robust plan with a better strategic plan. The generalized linear model with current values indicated that the best explanatory variables were the unemployment rate in the construction sector and the Brent oil price. SVM performed better than the GLM at with the current values variables and thus making a hint at the existence of the nonlinear relationship between the variables. However, adding the lagged values of the variable to the pool of the independent variables resulted in almost the same performance between the SVM and GLM. Meaning that a GLM model with lagged variables performed similarly to the SVM with the current values while adding the lagged values did not increase the performance of the SVM model. The multivariate model's performance is no better than other methods tried earlier by the authors, such as a univariate autoregressive moving average model [9] regressing on project frequency's past value. However, these multivariate models provide insight regarding the impact of environmental uncertainties on future project streams and thus could be valuable in long term strategic planning. Exploring other non-linear modeling techniques, such as neural networks for capturing more complicated relationships between the variables would be the next logical step in this research. The model developed in this study is limited to FDOT projects. However, new forecasting models specific for other databases can be built by following the same steps and adopting appropriate alternative sets of independent variables.

REFERENCES

[1] A. Shojaei, H. I. Moud, and I. Flood, "Forecasting Transportation Project Frequency using Multivariate Regression with Elastic Net Regularization Forecasting Transportation Project Frequency using Multivariate Regression with Elastic Net Regularization," in INFOCOMP 2018, The Eighth International Conference on Advanced Communications and Computation, 2018, no. July, pp. 74–79.

[2] M. Martinsuo, "Project portfolio management in practice and in context," Int. J. Proj. Manag., vol. 31, no. 6, pp. 794–803, Aug. 2013.

[3] A. Brown, J. Hinks, and J. Sneddon, "The facilities management role in new building procurement," Facilities, vol. 19, no. 3/4, pp. 119–130, 2001.

[4] World Bank, "Survey of International Construction Projec," 1996.

[5] R. G. Cooper, S. J. Edgett, and E. J. Kleinschmidt, "Portfolio management in new products: Lessons from the leaders, Part 1," Res. Technol. Manag., vol. 40, no. 5, pp. 16–28, 1997.

[6] "Project Portfolio Management Defined | Planview." [Online]. Available:

https://www.planview.com/resources/articles/project-portfolio-management-defined/. [Accessed: 01-Sep-2019].

[7] F. J. Fabozzi, H. M. Markowitz, P. N. Kolm, and F. Gupta, "Portfolio Selection," Theory Pract. Invest. Manag. Asset Alloc. Valuation, Portf. Constr. Strateg. Second Ed., vol. 7, no. 1, pp. 45–78, Mar. 2011.

[8] L. D. Dye and J. S. Pennypacker, "Project Portfolio Management and Managing Multiple Projects-Two Sides of the Same Coin?," in Managing Multiple Projects: Planning, Scheduling and Allocating Resources for Competitive Advantage, CRC Press, 2002, pp. 1–10.

[9] A. Shojaei and I. Flood, "Stochastic forecasting of project streams for construction project portfolio management," Vis. Eng., vol. 5, no. 1, p. 11, 2017.

[10] R. R. Trippi and J. K. Lee, Artificial intelligence in finance & investing : state-of-the-art technologies for securities selection and portfolio management. McGraw-Hill, Inc., 1996.

[11] A. D. Henriksen and A. J. Traynor, "A practical r&d project-selection scoring tool," IEEE Trans. Eng. Manag., vol. 46, no. 2, pp. 158–170, May 1999.

[12] J. A. Araúzo, J. Pajares, and A. Lopez-Paredes, "Simulating the dynamic scheduling of project portfolios," Simul. Model. Pract. Theory, vol. 18, no. 10, pp. 1428–1441, Nov. 2010.

[13] A. F. Carazo, T. Gómez, J. Molina, A. G. Hernández-Díaz, F. M. Guerrero, and R. Caballero, "Solving a comprehensive model for multiobjective project portfolio selection," Comput. Oper. Res., vol. 37, no. 4, pp. 630–639, Apr. 2010.

[14] F. Costantino, G. Di Gravio, and F. Nonino, "Project selection in project portfolio management: An artificial neural network model based on critical success factors," Int. J. Proj. Manag., vol. 33, no. 8, pp. 1744–1754, 2015.

[15] R. Demirer, R. R. Mau, and C. Shenoy, "Bayesian Networks: A Decision Tool to Improve Portfolio Risk Analysis.," J. Appl. Financ., vol. 16, no. 2, pp. 106–119, 2006.

[16] A. Shojaei and I. Flood, "Extending the Portfolio and Strategic Planning Horizon by Stochastic Forecasting of Unknown Future Projects," in The Seventh International Conference on Advanced Communications and Computation, INFOCOMP 2017, 2017, no. c, pp. 64–69.

[17] A. Shojaei and I. Flood, "Stochastic Forecasting of Unknown Future Project Streams for Strategic Portfolio Planning," in Computing in Civil Engineering 2017, 2017, pp. 280–288.

HW/SW Co-Design Approach to Optimize Embedded Systems on Reliability

Andreas Strasser, Philipp Stelzer, Christian Steger

Institute of Technical Informatics
Graz University of Technology
Graz, Austria
Email: {strasser, stelzer, steger}@tugraz.at

Norbert Druml

Infineon Technologies Austria AG
Graz, Austria
Email: norbert.druml@infineon.com

***Abstract*—Autonomous driving is disruptively changing the automotive industry. The importance of safety, reliability, and fault-tolerance is steadily increasing through the complexity and autonomy of self-driving cars. In the past, developers relied on the driver as a fail-safe backup to transfer the control and the responsibility to him in case of unexpected faults. In fully autonomous vehicles this backup solution will be not available anymore. This requires novel safety concepts and methodologies such as an optimization of high reliability of the systems. For optimization it is necessary to quantify different algorithm solutions from a safety point of view because this enables the possibility of comparing different solutions. In this publication, we are analyzing the consequences of different hardware and software algorithm implementations on component reliability. For this purpose we have designed two novel algorithm safety validation methodologies that allow the quantification of algorithms from a safety point of view and applied them to two independent use cases to evaluate the effects on component reliability. Both methodologies are used for optimizing the reliability of safety-critical automotive embedded systems for autonomous driving during Hardware/Software Co-Design.**

Keywords–Safety critical systems; Aging of circuits and systems; Safety Validation HW/SW; Failure-in-Time Analysis; Algorithm Safety Evaluation

I. INTRODUCTION

50 years ago started the future about fully autonomous driving. In the 1960s, Continental tested their driver-less car in the Contidrom in Germany. It was used as a prototype for tire testing to ensure constant testing conditions [2]. Nowadays, 50 years later this vision still exists in our society and Tesla has shown that autonomous driving is possible with their "Autopilot" [3]. Tesla has triggered the hype about autonomous driving and has pushed the society into a new era. This new era is changing the individual's daily routines about mobility and enables smart mobility.

Smart mobility will create a fully connected urban environment and will bring benefits to cities, better quality of life, reduced costs and more efficient energy usage [4]. To achieve the goal of autonomous driving and smart mobility, novel Advanced Driver-Assistance Systems (ADAS) are necessary. The two best-known ADAS are the Electronic Stability Control and the Anti-Lock Braking System, especially for their positive effect on active safety. Moreover, in the last years, a new generation of ADAS such as the Adaptive Cruise Control (ACC) has been established in middle class cars to avoid collisions. The next big step is introducing a comprehensive system enabling the perception of urban environment, which is one of the main goals of the PRYSTINE project [5].

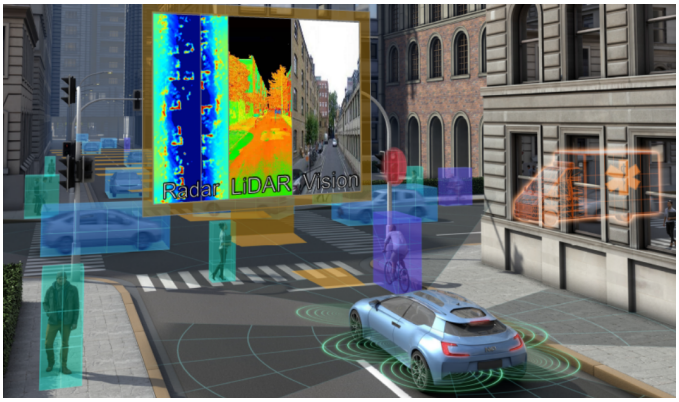


Figure 1. PRYSTINE's concept view of a fail-operational urban surround perception system [5].

PRYSTINE stands for Programmable Systems for Intelligence in Automobiles and is based on robust Radar and LiDAR sensor fusion to enable safe automated driving in urban and rural environments, as seen in Figure 1. These devices must be reliable, safe, and fail-operational to handle safety-critical situations independently [5].

In the past, developers of safety-critical automotive systems generally integrated the driver as the last safety chain link by handling over the control and the responsibility to the driver in unexpected situations or conditions. For fully autonomous vehicles, this fail-safe backup will not be available anymore because these vehicles needs to manage all critical unexpected situations on their own. This requires a rethinking of traditional safety concepts and methodologies. Novel safety-critical automotive embedded systems that will be equipped into autonomous vehicles needs to be high reliable, robust, and fail-operational [5]. One possibility that have been neglected in the past is about optimizing current systems from a safety point of view as increasing the component reliability. For this purpose, novel safety methodologies need to be developed that focus on optimizing embedded systems from a safety point of view.

This publication is an extended Version of the "FITness Assessment-Hardware Algorithm Safety Validation" [1] publication that was presented at the Ninth International Conference on Performance, Safety and Robustness in Complex Systems and Applications.

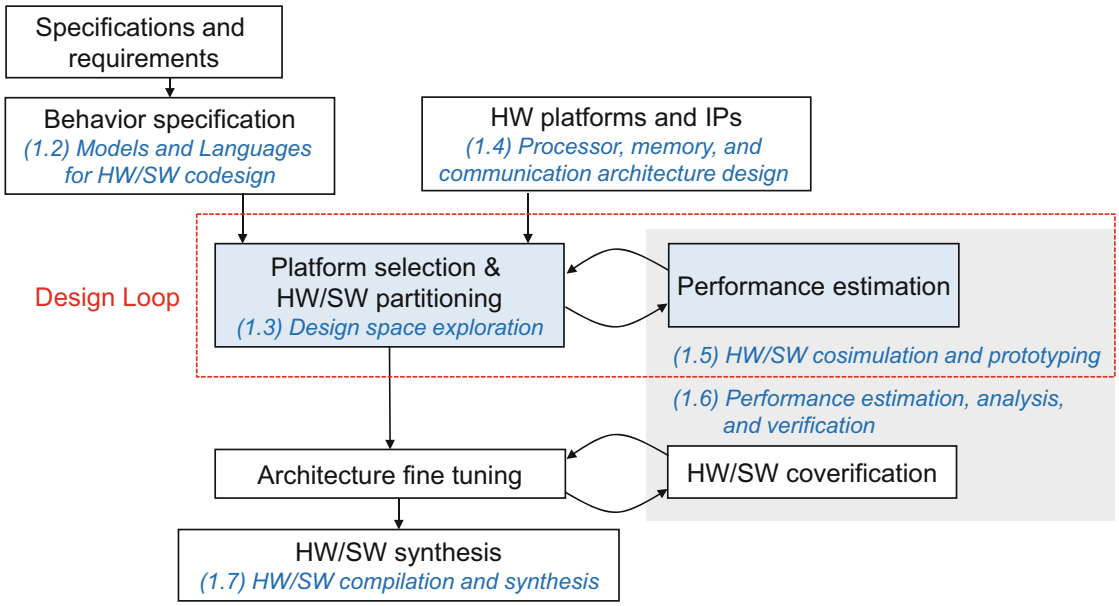


Figure 2. Overview of the HW/SW Co-Design design flow [6].

For this purpose, we will elaborate on the following two research questions:

- How can different hardware language description algorithm implementations be validated from a safety point of view?
- How can different software algorithm implementations be validated from a safety point of view?

The remainder of the paper is structured as follows. Related work will be provided in Section II. The method will be described in detail in Section IV and the results including a short discussion will be provided in Section VI. A summary of the findings will conclude this paper in Section VII.

II. Related Work

This section describes the related work in the field of component reliability considering HW/SW Co-design methodologies, software safety, hardware safety and component reliability.

A. Reliability Focused HW/SW Co-Design Methodologies

Schaumont [7] defines that the HW/SW Co-Design that is depicted in Figure 2 is used to design hardware and software components in a single design effort considering the partitioning and design of an application in terms of fixed and flexible components. In general, the most driving factor for the usage of the HW/SW Co-Design methodology is about making trade-offs, as depicted in Figure 3, between conflicting objectives such as performance, energy efficiency through fixed hardware implementations and flexibility through the usage of software implementations [7].

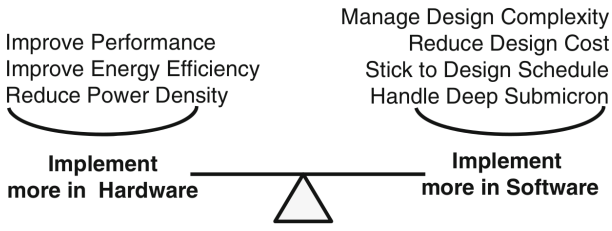


Figure 3. HW/SW Co-Design driving factors [7].

Beside the most common driving factors such as energy and performance there are also other factors that are more important in other domains such as the reliability for safety-critical embedded systems. Vargas et al. [8] introduced a novel HW/SW Co-Design approach that focus on the reliability of the overall system. Their approach decides on the basis of system reliability requirements which parts are partitioned into hardware or software including a verification of the overall reliability of the system. Vargas et al. focused in their publication on the correct function of the overall system and introduced primary hardware redundancy. Another work is the publication of Tosun et al. [9] that focus on soft errors such as bit flips. Both frameworks clearly shows that the overall reliability of safety-critical embedded systems are able to be improved by specific HW/SW Co-Design approaches. Nevertheless, both frameworks do not consider the component reliability of the hardware parts that are measured as the Failure in Time (FIT) Rate.

B. Software Design for Functional Safety

Nancy Leveson is one of the most known safety specialists and have published a book about software safety [10]. In 1995 Leveson described that in general software developers threat the computer as a stimulus-response system and that they seldom look beyond the computer. Consequently, software engineers usually constructed software without thinking about effects of the software on system safety [10]. 23 years later the perception of safety-critical software engineering has been improved and engineers are aware about the influences of software on the overall safety level [11]–[15].

Leveson [10] describes two common methodologies to

ensure run-time safety of safety-critical software systems: Dynamic and Static Analysis. Dynamic Analysis is a detection method for software errors or functional errors during run-time. Static Analysis by contrast focuses on formal errors such as race conditions or buffer overflows. Nowadays these two techniques have been advanced to frameworks that enhance the validation process.

Cruickshank et al. [11] have introduced a novel validation metrics framework for validating software safety requirements and have applied the method on a fictitious safety-critical surface-to-air missile system. Cruickshank et al. described that their framework supported the early identification of potential safety problems [11]. Baudin et al. [16] have described their novel tool for safety validation called "CAVEAT". CAVEAT is a statistical analysis tool to verify safety critical software and is used by Airbus to validate pieces of code as early as possible [16]. Michael et al. [15] also introduced a novel Hazard Analysis and Validation Metrics Framework. This framework is able to gauge the sufficiency of software safety requirements in the early software development process [15]. These frameworks illustrate the need of advanced methodologies to support safety-critical software development. However, these frameworks do not consider a validation of different algorithm implementations on the affects of component reliability.

Software algorithm validation is widely used to compare different implementations with respect to power consumption or run-time. Rashid et al. [17] have compared different sorting algorithms that are implemented in different programming languages on mobile devices. Their results clearly show that different implementations results in different power consumptions. Another example is the analysis of energy consumption of sorting algorithms on smartphones of Verma et al. [18]. Verma et al. have found out that the energy consumption depends on the data size as well as on the implemented sorting algorithm [18]. Bunse et al. have explored the energy consumption of data sorting algorithms in embedded environments and in their work different algorithms resulted in different power consumption. According to the automotive functional safety standard "ISO 26262" [19] power consumption is related to component reliability.

C. Hardware Design for Functional Safety

The validation of algorithms is an important method for achieving certain requirements such as area, power dissipation or run time. Therefore, there are numerous articles about enhancing efficiency of fault-tolerant mechanisms through algorithm substitution [20] [21] [22]. Rossi et al. analyze the power consumption of fault-tolerant buses by comparing different Hamming code implementations with their novel Dual Rail coding scheme [20]. Also, Nayak et al. emphasize the low power dissipation of their novel Hamming code components [21]. Another example is the work of Shao et al. about power dissipation comparison between the novel adaptive pre-processing approach for convolution codes of Viterbi decoders with conventional decoders [22]. Khezripour et al. provide another example for validating different fault-tolerant multi processor architectures by power dissipation [23]. Unfortunately, power dissipation is just one factor for reliability of safety-critical components and insufficient for safety validation.

The most important indicator for safety at hardware level is the component reliability, which is measured in failure in time (FIT) rates [19]. Component reliability is the main indicator for safe hardware components and describes the quantity of failures in a specific time interval, mostly one billion hours [19]. These values can be calculated by specific standards for electronic component reliability such as the IEC TR 62380 [24] or statistically collected by field tests. Oftentimes, these field test have already been conducted by the manufacturers and are compiled in specific data-sheets for component reliability [25]. For each component, the data-sheets usually contain the specific FIT Rate for a certain temperature. To determine the FIT Rate for other temperatures, the Arrhenius equation as seen in (1) can be used.

$$DF = e^{\frac{E_a}{k} \cdot (\frac{1}{T_{use}} - \frac{1}{T_{stress}}))} \quad (1)$$

where:

DF	De-rating Factor
E_a	Activation Energy in eV
k	Boltzmann Constant (8.167303 x 10^{-5} ev/K)
T_{use}	Use Junction Temperature in K
T_{stress}	Stress Junction Temperature in K

The Arrhenius Equation requires the Junction Temperature instead of Temperature values. The Junction Temperature represents the highest operation temperature of the semiconductor and considers the Ambient Temperature, Thermal Resistance of the package as well as the Power Dissipation as seen in (2).

$$T_j = T_{amb} + P_{dis} \cdot \theta_{ja} \quad (2)$$

where:

T_{amb}	Ambient Temperature
P_{dis}	Power Dissipation
θ_{ja}	Package Thermal Resistance Value

III. Problem Statement

The validation of different algorithms is crucial for designers to optimize their systems in terms of component reliability for highly robust and safe autonomous vehicles. Designers of safety-critical embedded systems should be able to pick the most safe algorithm with the advantage of lower FIT Rates. Especially for automotive Tier-1 companies lower FIT Rates imply higher component reliability, which is crucial for the economic success or failure of the whole system as profit margins are that small that every defect matters. Therefore, to support designers of safety-critical embedded systems, this publication's contributions to existing research are:

1) Developing novel methods for safety validation of hardware and software algorithms that is based on the approved ISO 26262 2nd Edition methods.
2) Applying the novel methods to quantify the differences between different algorithm implementations from a safety point of view.

IV. Component Reliability Focused HW/SW Co-Design Methodology

This section introduces two novel design processes that support designers of safety-critical embedded systems to find the most reliable solution during the HW/SW Co-Design

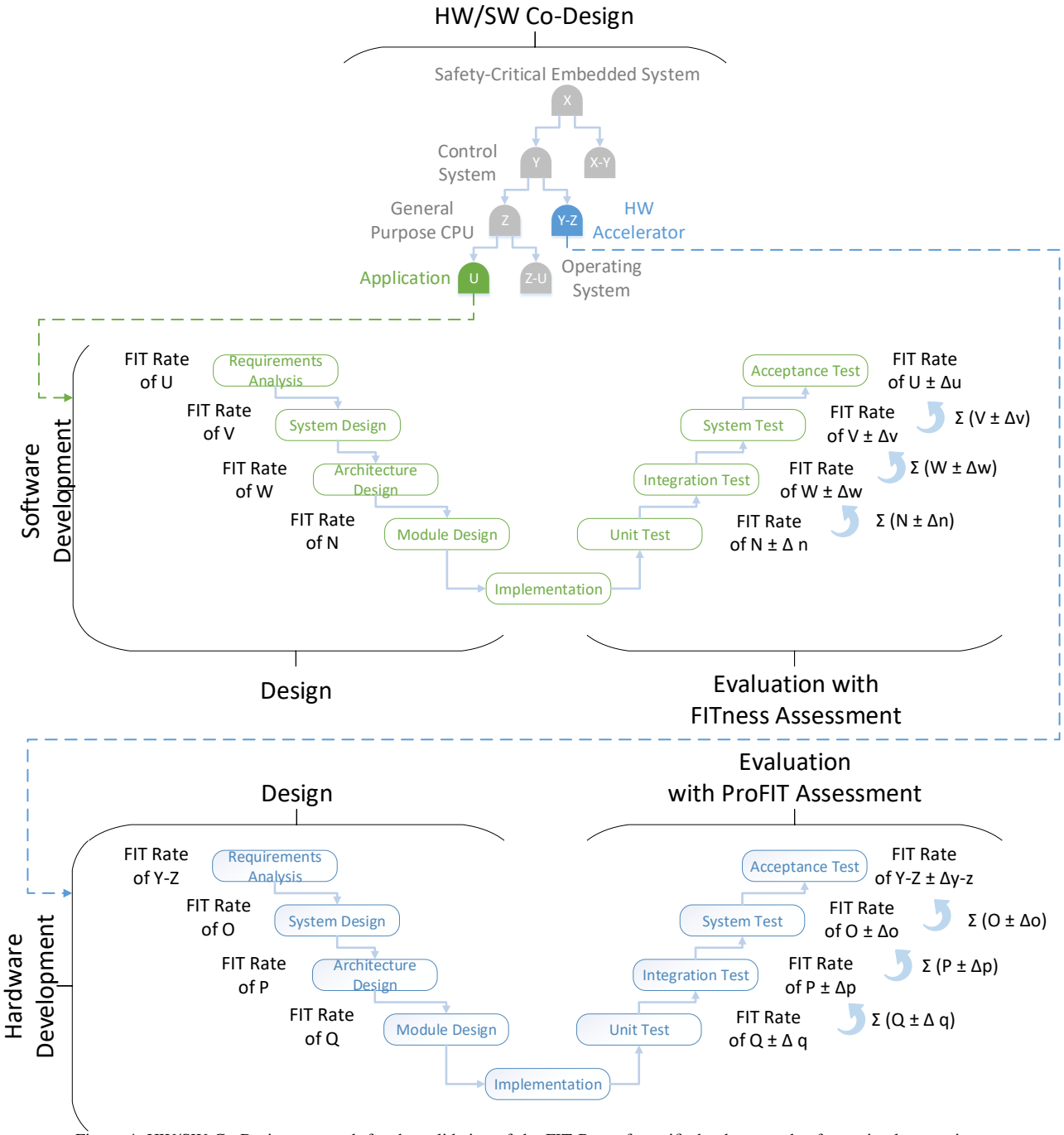


Figure 4. HW/SW Co-Design approach for the validation of the FIT Rate of specific hardware and software implementations.

process. The most reliable solution in this case is defined as the system with the lowest FIT Rate. To compare different hardware and software solutions it is necessary to measure the specific FIT Rate of each algorithm implementation. For this purpose, we need to introduce two novel measurement methodologies that enable the FIT Rate measurement. These two measurement methodologies that are presented in this publication are:

- **FITness Assessment - Hardware Reliability Evaluation** The "FITness Assessment" approach enables the FIT Rate determination of algorithms that are implemented in hardware description languages such as VHDL.
- **ProFIT Assessment - Software Reliability Evaluation** The "ProFIT Assessment" approach evaluates the FIT Rate of software implemented algorithms that are executed on micro-controller.

The FITness Assessment focuses on the estimation and validation of hardware related implementations and the ProFIT Assessment on software implementations. Both methods can easily be integrated in common HW/SW Co-Design design flows as depicted in Figure 4.

The novel HW/SW Co-Design approach that is enabled

through our two novel FIT Rate measurement approaches allows the evaluation of the FIT Rate of specific functionalities that are implemented in hardware or software. On the left side, a tree diagram of the overall safety-critical embedded system can be seen. The top leaf of the tree structure represents the whole embedded system and contains a FIT Rate of X. In the next hierarchical level the FIT Rate X is separated in the control system part and the additional hardware part that are represented with a FIT Rate of Y and X-Y. This strategy can be continued until we reach the smallest part of the overall system such as algorithms in software or hardware components. Based on this FIT Rate separation each designer and programmer is able to mind the overall FIT Rate of the system by complying with the given FIT Rate. Any deviance of a software algorithm can easily be recognized in the early phase of development and enables an intervention of the project team.

After the separation, each software programmer and hardware designer is able to determine if their solution matches the requirements of the designer considering the FIT Rate. Especially, the division of the overall FIT Rate into smaller sub-parts enables a reliability focused hardware-software development. A comparison between the designed reliability and the indeed reliability is possible through the summarization of the individual FIT Rates to the overall system. For this purpose, the individual FIT Rates of the software and hardware units are summed up to an overall system FIT Rate.

To enable this novel HW/SW Co-Design approach it is necessary to measure the FIT Rate of specific hardware and software implementations and this could be achieved by our novel hardware and software reliability evaluations called "FITness Assessment" and "ProFIT Assessment".

A. FITness Assessment - Hardware Reliability Evaluation

To validate different algorithms that are implemented in hardware description languages such as VHDL or Verilog, it is necessary to quantify the essential values. Based on the functional safety standard ISO 26262 2nd Edition's approved methods, the FIT Rate is the most important factor for safety-critical hardware components. As stated in the Related Work Section II, the De-rating Factor influences the FIT Rate and is expressed in the Arrhenius equation (1). Combined with the Temperature Junction equation it is obvious that the power dissipation is the most significant quantity that can be influenced by designers of digital circuits (see (3)).

$$DF = e^{\frac{E_a}{k} \cdot (\frac{1}{T_{use}} - \frac{1}{T_{amb} + P_{dis} \cdot \theta_{ja}})} \quad (3)$$

Consequently, by decreasing Power Dissipation the designer increases component reliability. For Field Programmable Gate Array (FPGA), the power dissipation primarily depends on static and dynamic power consumption. Based on these physical principles, our novel method FITness Assessment for algorithm safety validation on FPGAs is segmented in the following parts, as seen in Figure 5:

1) **Algorithm Implementation**
To guarantee similar conditions for different algorithms, it is necessary to implement a generic framework that allows implementing algorithms without major changes.

2) **Power Consumption Measurement**
For each algorithm, a particular measurement is recorded. It is advisable to record the generic framework without any algorithm to be able to determine the algorithms' power consumption by subtraction.

3) **Determination of Base FIT Rate**
The Base FIT Rate may be calculated by using the IEC TR 62380 [24] standard or analyzed statistically by field tests. Oftentimes, these field test have already been conducted by the manufacturers and are compiled in specific data-sheets for component reliability.

4) **De-rating Factor Calculation**
The De-rating Factor can be calculated with the Arrhenius equation and the related Thermal Junction equation as seen in (1) and (2).

5) **Identification of Effective FIT Rate**
The Effective FIT Rate reflects the Base FIT Rate for a specific temperature and can be calculated with:

$$FIT_{ef} = FIT_{base} \cdot DF \quad (4)$$

where:

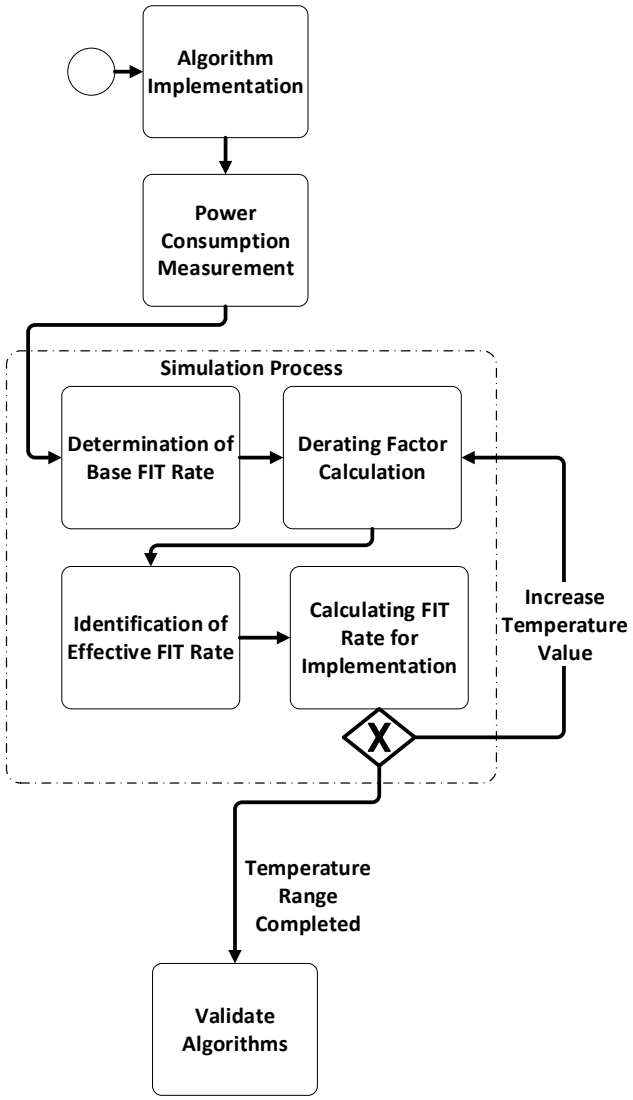


Figure 5. Workflow overview of our novel method FITness Assessment for algorithm validation from a safety point of view in Business Process Model and Notation.

FIT_{base} Base FIT Rate from FPGA Reliability Data-sheet
DF De-rating Factor as seen in (1)

6) **Calculating FIT Rate of the Implementation**
The Effective FIT Rate as seen in (4) represents the component reliability for the whole FPGA. However, an FPGA is made up of many different logic elements. Consequently, the Effective FIT Rate can be broken down into the amount used by each logical element as seen in (5).

$$FIT_{imp} = \frac{FIT_{ef}}{N_{le}} \quad (5)$$

where:
FIT_{ef} Effective FIT Rate as seen in (4)
N_{le} Total Number of Logic Elements of the specific FPGA taken out from Data-sheet

7) **Validate Algorithms**
The resulting FIT Rate of the implementation represents the FIT Rate of the specific algorithm and can be used for validation. It is advisable to measure each algorithm once at room temperature conditions and simulate the rest of the temperature range by starting with the De-rating Factor Calculation.

B. ProFIT Assessment - Software Reliability Evaluation

Validating software algorithms for safety-critical systems from a safety point of view can be obtained by using our novel "ProFIT Assessment". This method enables the impact measurement of different software algorithm implementation on component reliability. Our novel method is using approved methods from the functional safety standard ISO 26262 2nd Edition [19] of the automotive industry. As a starting base we have used equation (3). This equation represents the impacts on the component FIT Rate as a function of the power consumption. In Related Work we have introduced scientific results that clearly shows that different software algorithm implementations results in different power consumption. Therefore, the De-rating Factor can be used to determine the specific software algorithm FIT Rate. Our "ProFIT Assessment" is using these relations and can be separated into five parts:

1) **Implementation**
Different algorithms will be implemented in software. For better results and accuracy it is advisable to implement a general framework where the algorithms can be exchanged without any major changes. The framework will be compiled and programmed onto a specific micro-controller. In general any micro-controller can be used but it is advisable to look for public available component reliability data-sheets.

2) **Measurement**
In this step the software algorithms will be run on micro-controller and the power dissipation is recorded. This step will be repeated for each implementation. As an output result a measurement report is created, which contains the measurement setup, the used micro-controller, software algorithm implementation, power consumption and ambient testing temperature. These details are necessary for further analysis.

3) **Calculating FIT**
The idea behind this step is that each software algorithm needs a specific amount of time and the power consumption is measured at a specific sampling rate. For each sample we are calculating the specific Base FIT Rate and relates it to the sampling duration. Summing up all the individual FIT Rates of each timeslice results in the specific FIT Rate of the software algorithm implementation for a specific temperature. The impacts of the different implementations over the whole temperature range will be determined through the simulation process afterwards.

a) **Junction Temperature**
At first we are calculating the specific Junction Temperature for the ambient testing temperature as seen in (2).

b) **De-rating Factor**
Secondly the specific De-rating Factor is determined with the Arrhenius equation as seen in (1).

c) **Base FIT Rate**
The base FIT Rate can be determined by multiplying the base FIT Rate from compo-

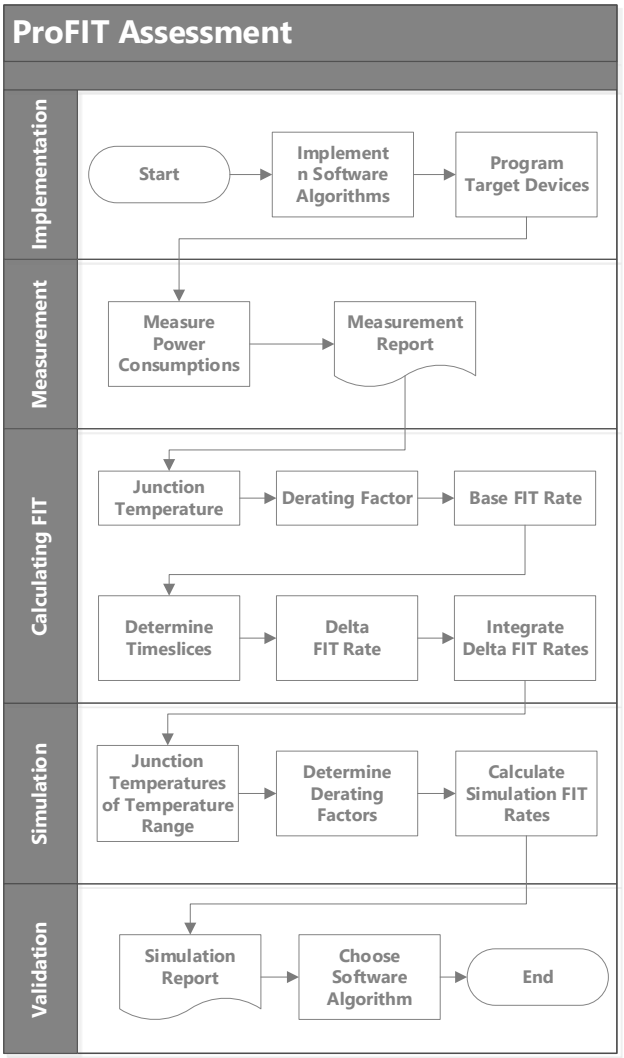


Figure 6. Work flow overview of our novel "ProFIT Assessment" method for software algorithm validation from a safety point of view.

nent reliability data-sheet with the De-rating Factor.

$$FIT_{Base} = DF \cdot FIT_{Ds} \quad (6)$$

where:

DF De-rating Factor as seen in (1)
FIT_{Ds} Base FIT Rate of Component Reliability Data-sheet

d) **Determine Time-slices**
In this step the Base FIT Rate will be adapted to the specific run-time.

$$FIT_{Timeslice} = FIT_{Base} \cdot \frac{T_{Sampling}}{T_{Runtime}} \quad (7)$$

where:

FIT_{Base} Base FIT Rate as seen in (6)
$T_{Sampling}$ Measurement Sampling Time
$T_{Run\text{-}time}$ Run-time of the Measurement

e) **Integrate Delta FIT Rates**
To determine the Software FIT Rate it is necessary to accumulate all individual Time-slices.

$$FIT_{Algorithm} = \sum_{1}^{n} FIT_{Ts} \quad (8)$$

$$n = \frac{T_{Runtime}}{T_{SamplingRate}} \quad (9)$$

where:

FIT_{Ts} Time-slice FIT Rate as seen in (7)
$T_{Sampling}$ Measurement Sampling Time
$T_{Run\text{-}time}$ Run-time of the Measurement

4) **Simulation**
The simulation step is necessary to determine the software algorithm FIT Ratio over the whole operational temperature range. The power consumption variation will be neglected because it affects all algorithm implementations equally.

a) **Junction Temperatures of Temperature Range**
This step is similar as during the Calculating FIT Rate step except the use of the whole operational temperature range.

b) **Determine De-rating Factors** This step is equal as seen in (1).

c) **Calculate Simulation FIT Rates**
This step is equal as seen in (6).

5) **Validation**
After the simulation there will be a Simulation Report with the specific FIT Rates for the whole operational temperature range. This can be used as a decision support to pick the most reliable software algorithm implementation.

V. Test Setup

This section describes the practical results of this publication by introducing the testing environment and the final results of the experiments. The validation of the HW/SW Co-design approach was divided in a software and hardware part and both parts have been validated independently.

A. FITness Assessment Evaluation Setup

In our research question, we analyze the differences between Single Error Correction - Double Error Detection (SEC-DED) and Double Error Correction (DEC). For this purpose, we chose the Hamming code for SEC-DED as this code is recommended in the new ISO 26262 2nd Edition and the BCH-code for DEC, especially because other ECC algorithms are often based on this concept and both algorithms fulfill the following requirements:

- 32 Bit data size
- Combinatorical Logic
- Including Fault Injection Module
- SEC-DED or DEC Functionality

The generic algorithm framework contains a test-bench with an automatic up-counter as well as a validator (see Figure 8). Both algorithms can be exchanged in the framework without any major changes. This enables a precise validation from a safety point of view.

In our test setup, we use the MAX1000 - IoT Maker Board by Trenz Electronic. This device is a small maker board for prototyping with sparse additional components. The main controller is the MAX10 10M08SAU169C8G, an FPGA device by Intel. For our research, the main advantages of using this board are:

- Small amount of additional hardware components
- Availability of Reliability Data-sheet

This board also contains an FTDI chip that draws about 50 mA on average, which we will subtract out for our analysis. The power consumption measurement is performed by the Mobile Device Power Monitor of Monsoon Solutions. The big advantage of this power monitor is the direct measurement of USB devices. The entire measurement setup is shown in Figures 7 and 9 and contains the following software and hardware parts:

- Quartus Prime 18.0 (Intel)
- Power Tool 5.0.0.23 (Monsoon Solutions)
- Mobile Device Power Monitor (Monsoon Solutions)
- MAX1000 - IoT Maker Board (Trenz Electronic)

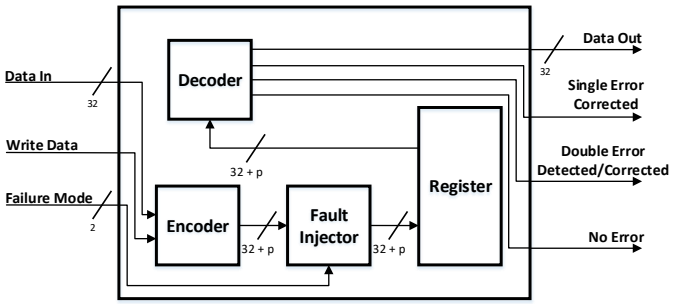


Figure 7. Pin configuration of both algorithms including an overview of functional blocks inside.

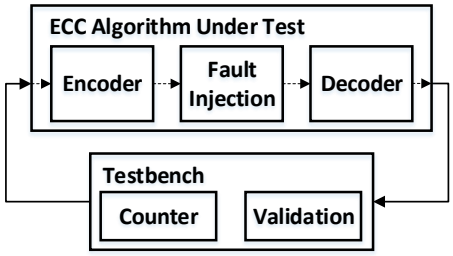


Figure 8. General framework for ECC algorithm validation including test-bench and ECC algorithm.

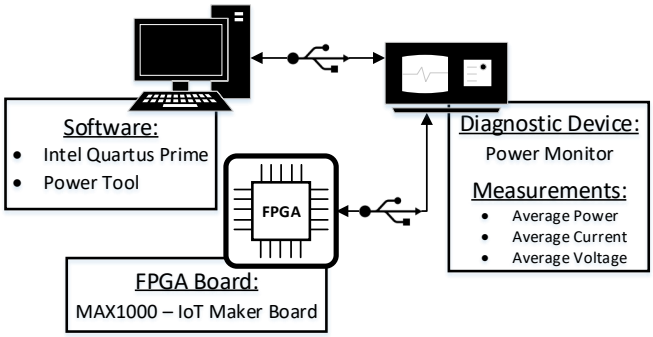


Figure 9. Overview of the entire measurement setup including software and hardware components.

B. ProFIT Assessment Evaluation Setup

For testing purpose we have chosen sorting algorithms as test candidates. The reasons for us are:

- Very often used
- Easy to understand
- Many different algorithms available
- Comparable results of power consumption available as seen in Section II-B

The sorting algorithms we chose are widely used and known and are known as:

- Binary Insertion Sort
- Heapsort
- Insertion Sort
- Mergesort
- Quicksort
- Shell Sort

All sorting algorithms were implemented in C programming language and programmed onto a micro-controller. For the micro-controller we have chosen the "MSP430 FR5969" from Texas Instruments by the following reasons:

- Measure Power Consumption with EnergyTrace++ Technology in "Code Composer Studio"
- Qualified for automotive usage
- Low-Power Device
- FIT Rates publicly available

As a operational temperature range for the simulation part we have chosen $-40°C$ up to $140°C$. This range is higher than the recommended operating conditions from the data-sheet but for our tests it is not relevant.
Test Setup Summary:

- Code Composer Studio 8.1
- MSP430 FR5969
- 6 different Sorting Algorithms
- 400 Numbers to Sort
- $-40°C$ up to $140°C$ Temperature Range for Simulation

VI. Results

A. FITness Assessment Evaluation

This section summarizes our results of the comparison of SEC-DED and DEC ECC algorithm. The validation was performed with our novel FITness Assessment method for algorithm validation from a safety point of view as described in Section IV.

The first algorithm we implemented was the Hamming code, which is a SEC-DED ECC algorithm. The implementation reserves 45 logic elements of the used FPGA and the whole board has an average power dissipation of 571.78 mW. With the second BCH-code DEC ECC algorithm, the board consumes an average of 599.05 mW and assigns 65 logic elements. The first result shows a difference between both algorithms in logic elements as well as in power dissipation resulting in a varying FIT Rate. The next step is the simulation process over the whole temperature range. We selected a temperature range between -40°C and 125°C and the values of Table I were used for the simulation process. In our simulation we neglected the alteration of power dissipation through temperature because it would affect both ECC implementations evenly.

Figure 10 points out that both algorithms vary in their FIT Rate and rise exponentially with increasing temperature. The FIT Rate may be neglected for temperatures up to 40 °C.

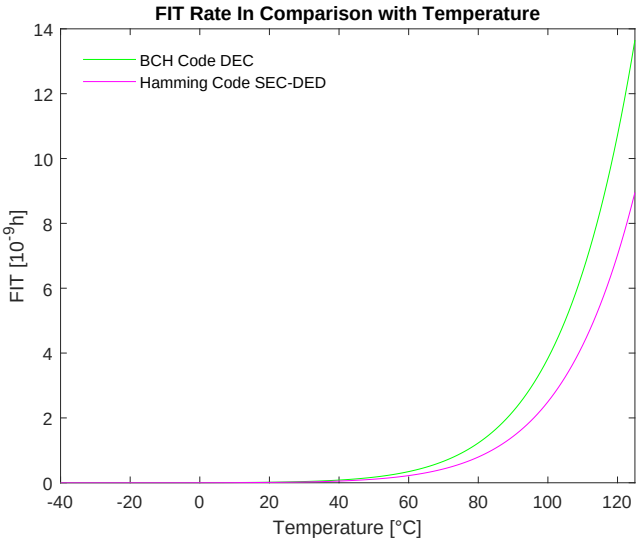


Figure 10. Simulation results of the resulted FIT Rates between -40°C and 125°C for both ECC implementations.

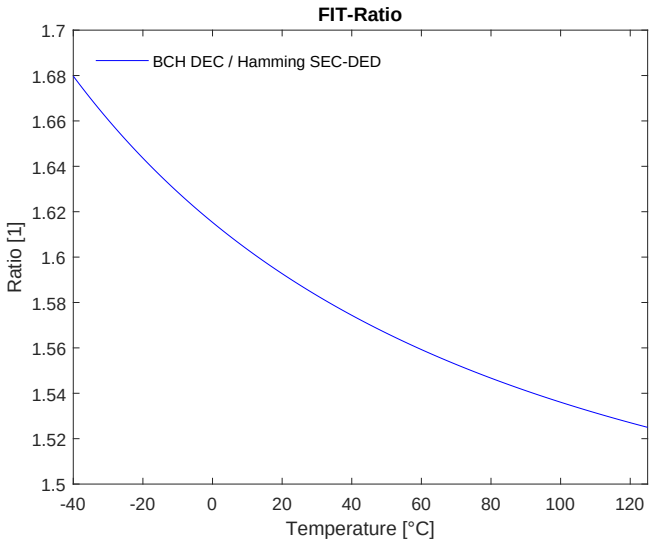


Figure 11. Overview of the FIT Rate overhead between SEC-DED and DEC ECC algorithm.

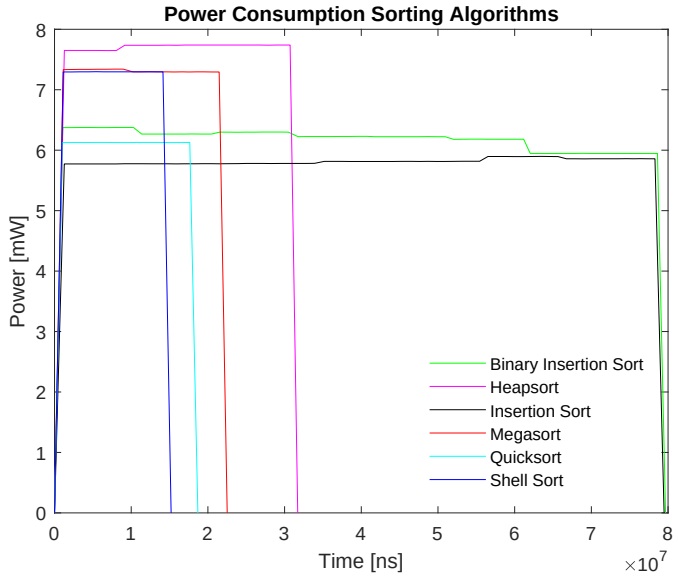


Figure 12. Power consumption results of the implemented sorting algorithms at 25°C ambient temperature.

TABLE I. RESULTS OF THE RESERVED LOGIC ELEMENTS AND AVERAGE TOTAL POWER DISSIPATION OF BOTH ECC IMPLEMENTATIONS.

	Hamming Code	BCH-Code
Used Logic Elements	45	65
Total Average Power Dissipation	571.78 mW	599.05 mW

The Hamming code with SEC-DED shows a better FIT Rate indicating more reliability of the hardware components which results in a higher safety level. The reason for this difference is the greater number of logic elements used for the DEC ECC algorithm and the resulting increase of power dissipation. The higher power dissipation results in a higher Thermal Junction temperature as seen in (2), which leads to a higher FIT Rate.

Both algorithms were implemented without any safety measures. This means that any damage to the Logic Element of the FPGA leads to failure of the whole ECC algorithm and the safe memory block. The ECC algorithm is the measure against SEU related altered flip flops inside the memory block, which decreases the specific FIT Rate of the memory block. The results of Figure 10 do not represent the FIT Rates of the memory block but the FIT Rate of the pure ECC implementation. It is important to understand that the ability of more bit error correction is not considered for the algorithm validation because it only positively influences the FIT Rate of the memory block.

Moreover, it is important to understand that the absolute values of the FIT Rate always correlate to a specific FPGA. Consequently, it is advantageous to look at the ratio between the algorithms because this gives a better overview of the overhead. The SEC-DED/DEC ECC FIT Ratio is depicted in Figure 11. The FIT Ratio overhead of the DEC ECC algorithm is slighly decreasing with increasing temperature, which is negligible in practice.

We recommend using the Hamming code algorithm for SEC-DED error correction for 32 bit memory size registers in automotive LiDAR systems. The SEC-DED algorithm used in our experiment resulted in a FIT Rate that was at least 52% lower than the DEC ECC algorithm.

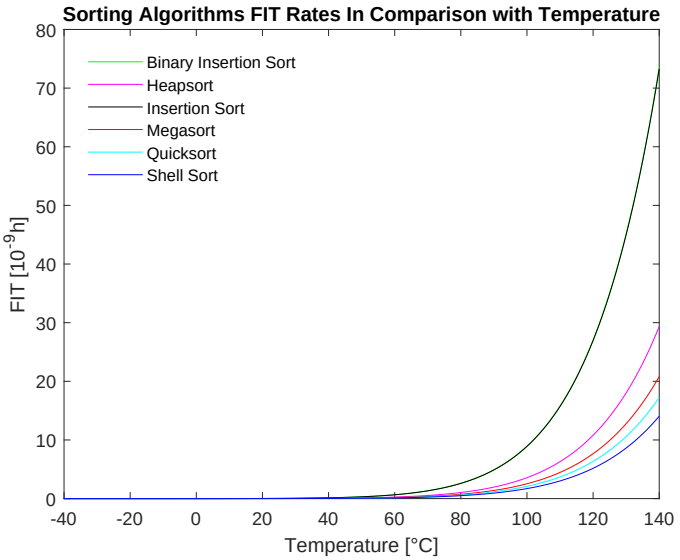


Figure 13. Simulation results of the sorting algorithms between −40°C and 140°C.

B. ProFIT Assessment Evaluation

In this section we are presenting our results of applying our novel "ProFIT Assessment" on sorting algorithms. This method enables the possibility to validate software algorithms from a safety point of view. It is important to understand that we are not comparing sorting algorithms instead we are applying our method on the sorting algorithms.

All algorithms are implemented in C and were tested on the "MSP430 FR5969" micro-controller board. This board has the possibility to measure the power consumption of each algorithm directly in the "Code Composer Studio". Table II gives an overview about our power measurement results of the implemented sorting algorithms. These algorithms were implemented in C and were executed on the "MSP430 FR5969" micro-controller board. The "Shell Sort" algorithm was in our test case the fastest at run-time and needed the least energy during run-time. Figure 12 shows the results of our power consumption measurements. In our setup "Shell Sort"

TABLE II. **Overview of the Power Consumption measurements of all C implemented sorting algorithms at $25°C$ ambient temperature.**

	Average Power in mA	Energy in uJ	Time in ms
Binary Insertion Sort	6.18	438.2	77.53
Heapsort	7.72	178.4	31.71
Insertion Sort	5.82	440.0	79.48
Mergesort	7.31	124.8	22.52
Quicksort	6.12	60.7	18.69
Shell Sort	7.30	58.5	15.20

TABLE III. **Results of the algorithm FIT Rates calculation of the implemented sorting algorithms on the MSP430 FR5969 micro-controller board.**

	FIT Rate in 10^{-9}
Binary Insertion Sort	1.87204922
Heapsort	0.747313371
Insertion Sort	1.865387949
Mergesort	0.529712728
Quicksort	0.438742916
Shell Sort	0.357627573

had the best run-time performance and "Binary Insertion Sort" had the worst run-time. This result clearly shows that different algorithm implementations result in different power consumptions. With these results the specific algorithm FIT Rates can be determined with the equations that have been introduced in IV-B.

The provided Table III represents the FIT Rate for a specific ambient temperature. In our case we have calculated the FIT Rate for the test ambient temperature of $25°C$. For other temperatures a simulation over the whole temperature range is necessary. For this purpose we have used the Arrhenius equation as seen in (1). In Figure 13 the FIT Rates of the implemented algorithms is displayed with the behavior over the whole temperature range. It can be seen that "Shell Sort" has the best FIT Rate over the whole temperature range and "Binary Insertion Sort" is the worst. For temperatures up to $50°C$ it does not matter what kind of algorithm is used but afterwards it has an affect on the component reliability and therefore on the overall safety level.

VII. Conclusion

In this publication, we introduced a novel HW/SW Co-Design approach that is optimizing the reliability of safety-critical automotive systems. To enable this approach, we have introduced two novel reliability evaluation methodologies that are able to analyze the impacts of different hardware and software algorithms on the component reliability also called Failure-In-Time Rate.

The hardware related part of the publication introduced the FITness Assessment, a novel component reliability hardware evaluation methodology and this was used to evaluate two different error correction code algorithms (SEC-DED and DEC ECC) from a safety perspective. The software related part introduced the ProFIT Assessment, a novel component reliability software evaluation methodology and this was used to analyze the impacts of six different sorting algorithms (Binary Insertion Sort, Heapsort, Insertion Sort, Mergesort, Quicksort and Shell Sort) to the overall component reliability of the micro-controller part of the overall embedded system.

Both methods are based on approved methods of the novel automotive functional safety standard ISO 26262 2nd Edition. The result clearly shows that different hardware and software algorithms lead to different FIT Rates.

FITness Assessment allowed the measurement of each algorithm's specific FIT Rate, facilitating the selection of the most reliable ECC algorithm. Our case shows a DEC-ECC algorithm that has a higher FIT Rate than the SEC-DED ECC algorithm.

ProFIT Assessment focuses on evaluating component reliability of software algorithms on micro-controllers. In our results we have showed that safety validation of software algorithms is possible and that different algorithm implementations can result in different component reliability. These differences should not be neglected because they have an impact from a safety point of view.

The FIT Rate reflects component reliability, which is an important hardware indicator for safety. These differences should not be neglected from a safety as well as from a business point of view. The FIT Rate also statistically indicates the amount of defective components, which is an economically important indicator as lower FIT rates also result in less defect components.

Fault-tolerance, safety and reliability will become more and more important in the next years because of autonomous driving. The novel introduced FITness Assessment enables the validation of different hardware algorithms to be able to select the most reliable one, which helps improve the overall safety level of the automotive vehicle by increasing component reliability. "ProFIT Assessment", the second method we introduced in this publication enables the possibility to validate the FIT Rate of software algorithm implementations and enables the possibility to choose the most reliable one. Both methodologies can be used for HW/SW Co-design for optimizing safety-critical automotive embedded systems from a safety point of view.

VIII. Acknowledgments

The authors would like to thank all national funding authorities and the ECSEL Joint Undertaking, which funded the PRYSTINE project under the grant agreement number 783190.

PRYSTINE is funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under the program "ICT of the Future" between May 2018 and April 2021 (grant number 865310). More information: https://iktderzukunft.at/en/.

References

[1] A. Strasser, P. Stelzer, C. Steger, and N. Druml, "FITness Assessment-Hardware Algorithm Safety Validation," in The Ninth International Conference on Performance, Safety and Robustness in Complex Systems and Applications, PESARO 2019, pp. 12–17.

[2] "50 Jahre fahrerloses Fahren: Pressematerial," 2014, URL: https://publicarea.admiralcloud.com/p/a49d3d8ba92f3cdbfa864f [accessed: 2019-11-11].

[3] M. Dikmen and C. Burns, "Trust in autonomous vehicles: The case of Tesla Autopilot and Summon," in 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct 2017, pp. 1093–1098.

[4] R. Faria, L. Brito, K. Baras, and J. Silva, "Smart mobility: A survey," in 2017 International Conference on Internet of Things for the Global Community (IoTGC), July 2017, pp. 1–8.

[5] N. Druml, G. Macher, M. Stolz, E. Armengaud, D. Watzenig, C. Steger, T. Herndl, A. Eckel, A. Ryabokon, A. Hoess, S. Kumar, G. Dimitrakopoulos, and H. Roedig, "PRYSTINE - PRogrammable sYSTems for INtelligence in AutomobilEs," in 2018 21st Euromicro Conference on Digital System Design (DSD), Aug 2018, pp. 618–626.

[6] S. Ha and J. Teich, Handbook of hardware/software codesign. Springer Publishing Company, Incorporated, 2017, ISBN: 978-94-017-7268-6.

[7] P. R. Schaumont, A practical introduction to hardware/software codesign. Springer Science & Business Media, 2012, ISBN: 978-1-4614-3736-9.

[8] F. Vargas, E. Bezerra, L. Wulff, and D. Barros, "Optimizing HW/SW codesign towards reliability for critical-application systems," in Proceedings Seventh Asian Test Symposium (ATS'98)(Cat. No. 98TB100259). IEEE, 1998, pp. 52–57.

[9] S. Tosun, N. Mansouri, E. Arvas, M. Kandemir, Y. Xie, and W.-L. Hung, "Reliability-centric hardware/software co-design," in Sixth international symposium on quality electronic design (isqed'05). IEEE, 2005, pp. 375–380.

[10] N. G. Leveson and J. Diaz-Herrera, Safeware: system safety and computers. Addison-Wesley Reading, 1995, vol. 680, ISBN: 978-0201119725.

[11] K. J. Cruickshank, J. B. Michael, and M. Shing, "A Validation Metrics Framework for safety-critical software-intensive Systems," in 2009 IEEE International Conference on System of Systems Engineering (SoSE), May 2009, pp. 1–8.

[12] W. Ahmad, U. Qamar, and S. Hassan, "Analyzing different validation and verification techniques for safety critical software systems," in 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Sept 2015, pp. 367–370.

[13] N. G. Leveson and P. R. Harvey, "Analyzing Software Safety," IEEE Transactions on Software Engineering, vol. SE-9, no. 5, Sept 1983, pp. 569–579.

[14] E. M. E. Koursi and G. Mariano, "Assessment and certification of safety critical software," in Proceedings of the 5th Biannual World Automation Congress, vol. 14, June 2002, pp. 51–57.

[15] J. B. Michael, M. Shing, K. J. Cruickshank, and P. J. Redmond, "Hazard Analysis and Validation Metrics Framework for System of Systems Software Safety," IEEE Systems Journal, vol. 4, no. 2, June 2010, pp. 186–197.

[16] P. Baudin, A. Pacalet, J. Raguideau, D. Schoen, and N. Williams, "Caveat: a tool for software validation," in Proceedings International Conference on Dependable Systems and Networks, June 2002, p. 537.

[17] M. Rashid, L. Ardito, and M. Torchiano, "Energy Consumption Analysis of Algorithms Implementations," in 2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Oct 2015, pp. 1–4.

[18] M. Verma and K. Chowdhary, "Analysis of Energy Consumption of Sorting Algorithms on Smartphones," in Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018, pp. 26–27.

[19] I. n. E. ISO, "Draft 26262 2nd Edition: Road vehicles-Functional safety," International Standard ISO/FDIS, vol. 26262, 2018.

[20] D. Rossi, A. K. Nieuwland, S. V. E. S. van Dijk, R. P. Kleihorst, and C. Metra, "Power Consumption of Fault Tolerant Busses," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 16, no. 5, May 2008, pp. 542–553.

[21] V. S. P. Nayak, C. Madhulika, and U. Pravali, "Design of low power hamming code encoding, decoding and correcting circuits using reversible logic," in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), May 2017, pp. 778–781.

[22] W. Shao and L. Brackenbury, "Pre-processing of convolutional codes for reducing decoding power consumption," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, March 2008, pp. 2957–2960.

[23] H. Khezripour and S. Pourmozaffari, "Fault Tolerance and Power Consumption Analysis on Chip-Multi Processors Architectures," in 2012 Seventh International Conference on Availability, Reliability and Security, Aug 2012, pp. 301–306.

[24] T. IEC, "Iec 62380," Reliability data handbook–universal model for reliability prediction of electronics components, PCBs and equipment (emerged from UTEC 80-810 or RDF 2000), 2004.

[25] "Intel Reliability Report," 2018, URL: https://www.intel.com/content/www/us/en/programmable/support/quality-and-reliability/reports-tools/reliability-report/rel-report.html [accessed: 2019-11-11].

Practice of Formalised Conceptual Knowledge Complements Realising Multi-disciplinary Knowledge Resources for Natural Sciences and Humanities

Claus-Peter Rückemann
Westfälische Wilhelms-Universität Münster (WWU), Germany;
Knowledge in Motion, DIMF, Germany;
Leibniz Universität Hannover, Germany
Email: ruckema@uni-muenster.de

***Abstract*—The results and insights in the practice of multi-disciplinary knowledge resources presented in this article are based on the research and developments conducted during the last years. Many multi-disciplinary and practical geo-spatial data and application solutions require to employ integrated resources from disciplines of natural sciences and humanities to exploit holistically complex knowledge scenarios. In many cases, data and algorithms as well as workflows have to be created and tackled individually. The goal of this paper is to illustrate examples of integrated knowledge based on the ongoing research to create sustainable and innovative resources ans a comprehensive tool base of conceptual knowledge in geo-spatial application scenarios and beyond, for arbitrary knowledge context in any media. The solution should be complementary to the commonly available geo-spatial features and should fulfill a range of further criteria, especially for a coherent system of knowledge, multi-disciplinary, and data-centric. The result should allow to create and refer to facetted knowledge focussed on geo-spatial scenarios. The paper presents a range of multi-disciplinary knowledge complements in their formalised common conceptual knowledge and the results of an implementation based on the fundamental methodology of superordinate knowledge. The resulting solution is targeting geo-spatial application scenarios and has been used for many practical implementations over more than three decades. The resulting comprehensive subset of conceptual knowledge reference divisions, which was created from this long-term research, is available and first published with the cited research.**

Keywords–Conceptual Knowledge Complements; Multi-disciplinary Knowledge Resources' Practice; Superordinate Knowledge Methodology; UDC; Advanced Data-centric Computing.

I. INTRODUCTION

This extended research is based on the practical subset of conceptual knowledge, which was presented at the GEOProcessing 2019 conference in Athens, Greece [1]. Responding to the major interests from the public discussion in Athens, this research goes beyond plain methods and the limited view of 'data' and illustrates the formalisation and conceptual knowledge complements, the fundaments and organisation of realising multi-disciplinary knowledge resources, based on the Principles of Superordinate Knowledge.

The motivation of this research is to show a representative compilation of components, which are integral components in complex implementation scenarios and which are under creation and development for significant periods of time.

This paper presents different types of multi-disciplinary Knowledge Resources from complementary discipline collections and resources in reference with geo-spatial disciplines. The paper especially discusses the practice of formalised conceptual knowledge complements created with multi-disciplinary collections, containers, and referenced resources. It is a truth universally acknowledged, that geo-spatial disciplines are specialised and very much concentrating on providing solutions and tools for spatial data. When it gets to more complex situations, then, spatial data based on numerical coordinate reference systems and domain only approaches may not be sufficient. This is, for example, the case when describing the target knowledge with mathematical spacial facets and dimensions is not sufficient.

Many information and context maybe lost when knowledge is handled as plain data and mapped to preexisting attributes and categories. This is the case when a more holistic and more fundamental approach should be considered. In practice, associating different objectives and intentions, systematic knowledge, and physical features with knowledge, from methodology to implementation and realisation, can provide valuable solutions. The principles of superordinate knowledge provide such fundaments, from methodology to realisation.

The resulting solution is a comprehensive subset of conceptual knowledge, which should be complementary to the commonly available geo-spatial topologies, taxonomies, and features and the multi-disciplinary context. In consequence, the means of describing spatial data, objects, entities, and context should be substantially extended.

The rest of this paper is organised as follows. Sections II and III introduce the state of the art and motivation and discuss previous work, components, and used resources. Sections IV and V discuss basic practical examples of conceptual knowledge formalisation and present representative examples of multi-disciplinary knowledge complements. Section VI presents the resulting conceptual knowledge solution. Sections VII and VIII evaluate the resulting subset, directly related implementations, research, development, and cases studies and summarise the lessons learned, conclusions, and future work.

II. State of the art and motivation

It is most beneficial to have universal means to describe and document knowledge over all complements of the Knowledge Resources and being able to support formalisation. For advanced applications, it is also beneficial to have means, which can deal as conceptual knowledge framework.

Formalisation is the process of creating a defined set of rules, allowing a formal system to infer theorems from axioms. Formalised conceptual knowledge complements can be created employing references to consistent conceptual knowledge, here illustrated by references to UDC and UDC concordances. Sustainable knowledge and resources management was successfully implemented for environmental information and computation [2] along with a basis for environmental management, the International Organization for Standardization (ISO) 14000 series containing standard recommendations for assessment, evaluation, life cycle analysis, communication, and auditing [3]. Complementary implementations were also successfully created for advanced mathematical-computational scenarios [4].

Geo-spatial practice is focussed on providing cartographic means for certain space and environment. Widely employed tools are Geoscientific Information Systems and Geographic Information Systems. Most of these tools use geo-referenced data in order to organise and reference information. Available topologies can also provide for the categorisation of geo-spatial entities. All together these means are very limited when seen in a larger context as required for many complex application scenarios. Regarding that, one of the major deficits is the lack of a consistent and holistic knowledge concept. The fundaments of terminology and understanding knowledge are layed out by Aristotle [5], being an essential part of 'Ethics' [6]. Information sciences can very much benefit from Aristotle's fundaments and a knowledge-centric approach [7] but for building holistic and sustainable solutions, supporting a modern definition of knowledge [8], they need to go beyond the available technology-based approaches and hypothesis [9] as analysed in Platon's Phaidon.

In sciences, observation is one of the most important fundamental tasks [10]. But, as John Burroughs expressed "There is nothing in which people differ in more than in their powers of observation. Some are only half alive to what is going on around them." [11]. Triggered by the results of a systems cases study, it is obvious that superordinate systematic principles [12] are still widely missing in practice and education. Making a distinction and creating interfaces between methods and the implementation applications [13], the results of this research are illustrated here along with the practical example of the Knowledge Mapping methodology [14] enabling the creation of new object and entity context environments, e.g., implementing methods for knowledge mining context. This motivating background allows to build methods for knowledge mapping on a general methodological fundament.

The Organisation for Economic Co-operation and Development (OECD) has published principles and guidelines for access to research data from public funding [15]. The principles and guidelines are meant to apply to research data that are gathered using public funds for the purposes of producing publicly accessible knowledge. In this context, the OECD especially addresses knowledge, re-use, and knowledge generated from re-use. The means to achieve such recommendations even for complex scenarios is to use the principles of Superordinate Knowledge, which integrate arbitrary knowledge over theory and practice. Core assembly elements of Superordinate Knowledge [12] are:

- Methodology.
- Implementation.
- Realisation.

Separation and integration of assemblies have proven beneficial for building solutions with different disciplines, different levels of expertise. Comprehensive focussed subsets of conceptual knowledge can also provide excellent modular and standardised complements for information systems component implementations, e.g., for environmental information management and computation [2]. The conceptual knowledge reference divisions presented here are the result from more than three decades of scientific research in information science and multi-disciplinary knowledge.

III. Components, formalisation, and design

There is a number of criteria, which are of major significance for advanced and complex scenarios, especially, the conceptual knowledge and the knowledge resources have to provide. The resulting solution should fulfill a range of criteria in order to provide a most sustainable, flexible fundament, e.g.:

- Covering a coherent system of knowledge, supporting universal knowledge.
- Consistent implementation, quasi-standardised.
- Providing facetted conceptual knowledge features.
- Multi-disciplinary knowledge spectrum.
- Features for multi-lingual implementation.
- Data-centric implementation / method.
- Extensible concept.

Therefore, these criteria should allow advanced features, for example:

- Documentation of data, objects, scenarios, concepts, algorithms,
- universal context of knowledge criteria for all kind of knowledge in any media,
- knowledge documentation,
- knowledge consistent integration of publications and research data,
- knowledge mining,
- wide range of flexible implementation potential,
- supporting workflow features and documentation.

The Knowledge Resources can embrace a wide range of different types of complements, e.g., collections, containers, and other referenced resources.

These complements can be organised individually, e.g., due to the fact that often each collection maybe a long-term or even open-end matter of development, e.g., for an individual disciplines' task, research council or business.

In consequence, the primary design strategies of the core Knowledge Resources require a focus on data-centricity. Therefore, a central goal are sustainable, portable structures, which can be kept vital and knowledge-consistent with future developments.

The use of multi-disciplinary and consistent conceptual knowledge frameworks also contributes to the sustainable creation of sustainable concordances and solutions. For example, advanced knowledge discovery and computing can be realised very forward-looking and efficient when based on Knowledge Resources, concordances, and classification [16].

According with these strategies and goals, the selection of conceptual frameworks and integration is essential in order to ensure that conceptual references can be seamlessly developed with the development of growing Knowledge Resources in flexible and sustainable ways. The components and feature selection for a practical, formalised implementation are described with the following passages.

For the implementation of case studies, the modules are built by support of a number of major components and resources, which can be used for a wide range of applications, e.g., creation of resources and extraction of entities. The facility for consistently describing knowledge is a valuable quality, especially conceptual knowledge, e.g., using the Universal Decimal Classification (UDC) [17]. The UDC can be used for consistently formalising universal conceptual knowledge in multi-disciplinary and multi-lingual context.

The UDC is the world's foremost document indexing language in the form of a multi-lingual classification scheme covering all fields of knowledge and constitutes a sophisticated indexing and retrieval tool. The UDC is designed for subject description and indexing of content of information resources irrespective of the carrier, form, format, and language. UDC is an analytico-synthetic and facetted classification. It uses a knowledge presentation based on disciplines, with synthetic features. UDC schedules are organised as a coherent system of knowledge with associative relationships and references between concepts and related fields. Therefore, the UDC represents a most flexible facetted classification system for all kinds of knowledge in any media. The UDC provides 70,000 subdivisions, in 50 languages, which provides more than 3 million entries and verbal descriptions. The UDC is up to now internationally used in 130 countries, for 150,000–200,000 document collections worldwide. The classification has shown up being especially important for complex, facetted, multi-disciplinary, and long-term classification, e.g., with Knowledge Resources. The UDC is the best publicly available implementation of conceptual knowledge to illustrate the width and depth of knowledge dimensions. The UDC allows an efficient and effective processing of knowledge data and provides facilities to obtain a universal and systematical view on classified objects. Operational areas include author-side content classifications and museum collections, e.g., with documentation of resources, library content, bibliographic purposes on publications and references, for digital and realia objects. The Knowledge Resources objects and entities can refer to any conceptual knowledge, e.g., main UDC-based classes, which for this publication are taken from the multi-lingual UDC summary [17] released by the UDC Consortium under a Creative Commons license [18]. Facets can be created with any auxiliary tables, e.g., auxiliaries of place and space, time, language, and form as well as general characteristics, e.g., properties, materials, relations, processes, and operations, persons and personal characteristics. Symbolism and meaning have significant value for any application in information science [19]. Object entities can be associated with symbolism, which can also be referred to conceptual knowledge [20]. Observation and experience are essential [21] with the cognition process and the required referencing.

Multi-disciplinary Knowledge Resources have been designed, created, and developed with various scenarios and realisations over the last decades. Figure 1 shows the complements diagram of core resources and their conceptual organisation and implementations. The major complements consist of application resources and components, knowledge resources, and originary resources and sources.

The Knowledge Resources are in focus of this research. They can contain all the relevant conceptual knowledge references for the complements. The Knowledge Resources are flanked by application resources and components, which are based on module implementations and program components. Associated implementations imply scripts, bytecode and executables, sources from various high level languages. Implementations range from individual developments to common third party components.

The originary resources and sources imply realia and reference targets, which range from objects in libraries and museums to in situ objects.

The central Knowledge Resources, which are discussed in the research, cover the complements of factual, conceptual, procedural, and metacognitive knowledge. For this research, we are presenting examples of different content implementations, especially collections, containers, and referenced resources.

There are no limitations for the conceptual knowledge referenced in collections, containers, and referenced resources. However, a standard multi-disciplinary, multi-lingual classification is the UDC, which can be consistently used with all different types of implementations.

The UDC is used for any conceptual description, providing arbitrary means, e.g., for facetted classification, conceptual inter-references between objects, object association based on verbal description.

IV. Basic practical examples of formalisation

A. Required conceptual knowledge features

Data and objects result from public, commonly available, and specialised Knowledge Resources. The Knowledge Resources are containing factual and conceptual knowledge as well as documentation and instances of procedural and metacognitive knowledge. These resources contain multi-disciplinary and multi-lingual data and context. UDC provides auxiliary signs [22], which represent kinds of standardised "operations". UDC allows the creation of facetted knowledge using these features. The conceptual knowledge in focus

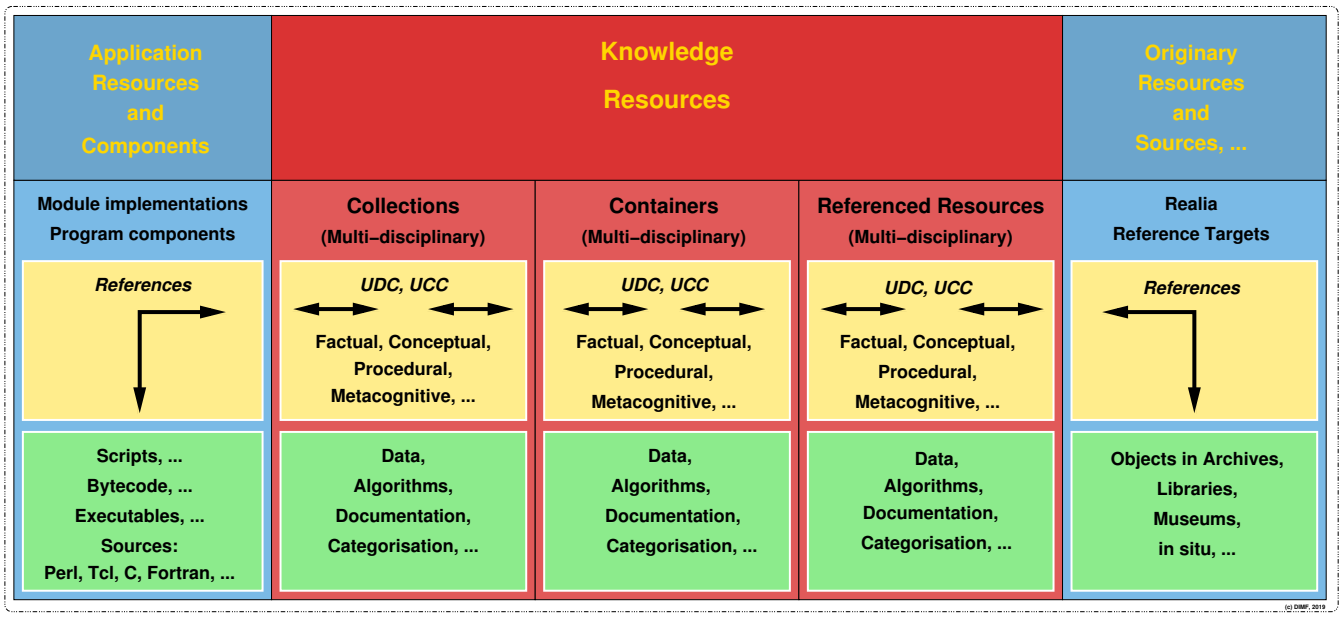


Figure 1. Multi-disciplinary Knowledge Resources: Complements diagram of core resources and their conceptual organisation and implementations. The three different types of Knowledge Resources discussed: Collections, containers, and referenced resources. Conceptual knowledge is illustrated via UDC and UCC.

requires to provide references to any universal knowledge context. References to UDC codes are capable to provide all the required context. The main tables provide an entry point to universal knowledge context [23]. For practical use, classification references can refer to UDC reference codes based on science and knowledge organisation [24]. For conceptual knowledge of place and spatial context the implementation requires to provide references to classification codes. The UDC provides references based on the common auxiliaries of place of the UDC [25]. In that context, besides universal knowledge, additional closely related references are required. UDC can provide appropriate references, e.g., geodesy, surveying, photogrammetry, remote sensing, cartography (UDC:528) [26] and geography, exploration, travel (UDC:910) [27], and nonliterary, nontextual representations of a region (UDC:912) [28].

B. Examples of conceptual knowledge application

Examples of conceptual knowledge reference divisions according with UDC (UDC:913, Regional geography, [29]; UDC:94, General history, [30]; UDC:(1/9), Common auxiliaries of place, [25]) and UDC conventions are shown in the following four small sample groups:

UDC:913(3) ⇒ Geography of the ancient world
UDC:913(3/9) ⇒ Geography of the individual regions and countries of the ancient and modern world

UDC:94(3) ⇒ History of the ancient world
UDC:94(3/9) ⇒ History of individual places of the ancient and modern world

UDC:94(37) ⇒ History of ancient Rome and Italy (to 5th century)
UDC:94(38) ⇒ History of ancient Greece

UDC:(37)(24) ⇒ Ancient Rome and Italy, below sea level
UDC:(38)(24) ⇒ Ancient Greece, below sea level

A little more complex facetted example, a single data object entity of a ship wreck realia as referred in a container of extended Knowledge Resources, is shown in Figure 2.

```
1  Lindos [Archaeology, Geophysics, Remote Sensing,
    Seafaring]:
2   Greek city, Rhodos Island, Dodekanese, Greece. ...
3   Object: Ship wreck.
4   Object-Type: Realia object.
5   Object-Location: 500\UD{m} SE of Hagios Pavlos Harbor.
6   %%IML: UDC:[902+903+...+904]+629.5+(38)+(4)+(24)...
7   %%IML: cite: YES 19810000 {LXK:Lindos; Rhodes; Ancient
     Greece; Archaeology; Artefacts; Ship wreck;} {UDC:...}
     {PAGE:--45..--58} LXCITE://Nikolitsis:1981:Rhodos
8   %%IML: ...
9   %%IML: OSMLocation: https://www.openstreetmap.org
     /...=36.08...%2C28.08...
10  %%IML: GoogleMapsLocation: http://maps.google.com/maps
     ...=...,...
```

Figure 2. Knowledge Resources, conceptual spatial and geo-references: Lindos object with ship wreck entity, Rhodes, Greece (excerpt).

Passages not relevant for demonstration and not adequate for privacy and safety reasons were shortened to ellipses. The object entity contains documentation, object categories and factual data, conceptual data references, a source reference [31], and data for geo-references. The conceptual knowledge comprises details of non geo-spatial domains, e.g., from main tables UDC:6 and UDC:9, and from geo-spatial context, e.g., auxiliary tables for place and space UDC:(24) UDC:(3/9). For this case, the object entity references can be resolved as:

UDC:902	⇒ Archaeology
UDC:903	⇒ Prehistory. Prehistoric remains, artefacts, antiquities
UDC:904	⇒ Cultural remains of historical times
UDC:629.5	⇒ Watercraft engineering. Marine engineering. Boats. Ships. Boatbuilding and shipbuilding
UDC:(38)	⇒ Ancient Greece
UDC:(4)	⇒ Europe
UDC:(24)	⇒ Below sea level. Underground. Subterranean

The references can hold further details and sub-contain additional information, e.g., UDC:903 further refers to artefacts in more detail. For a wider and deeper view, we have to refer to a number of successful projects, which were conducted by the author's group and various collaborators over the last decades. All these implementations are significantly based on the solution presented here.

V. IMPLEMENTATION OF KNOWLEDGE COMPLEMENTS

Example objects from the practical implementations of formalised conceptual knowledge complements for disciplines are shown in the following paragraphs. The examples show the range of contributing disciplines and the significance of features for a consistent conceptual knowledge implementation. Only excerpts of English language objects are shown for these examples. Anyhow the complements have to be available in multi-lingual instances of which the conceptual reference implementation must be able to handle. With the shown solution, the conceptual complements support about 50 languages.

A. Complementary environmental information

An example object from the environmental knowledge resources' complements is shown in Figure 3.

```
2004/35/EC [Environment, Climate, GIS, ...]:
 Directive 2004/35/EC, European Community, Environmental
  Liability Directive.
 %%SRC: 20050000 CPR
 The Environmental Liability Directive [ELD] 2004/35/EC
  is one of the most important instruments ...
 %%IML: UDC:502/504,551.581/551.582,341.1,(4)
 s. also ELD, EMS
```

Figure 3. Knowledge Resources, factual and conceptual complements: Environmental information, directives (excerpt).

The object carries factual knowledge and respective conceptual knowledge references. The full object refers to environmental information, climate information, and spatial context.

B. Complementary natural sciences

An example object from the natural sciences knowledge resources' complements is shown in Figure 4.

```
diffraction [Physics, Optics]:
 Deviation of a part of a ray due to the wave character
  of radiation.
 Diffraction occurs if rays hit the edge of an opaque
  obstacle.
 %%SRC: 1994, 2001, 2013 CPR
 %%IML: UDC-Object:535+535.42+550.3+550.8
 %%SRC: 2009 CPR
 %%IML: keyword-Context: KYW :: Physics, Optics, Waves,
  Seismics, Geophysics, Resolution, Exploration, Earth
  crust, Earth surface, Applied Geosciences, Archaeology,
   Scientific Computing, Data Processing, Statistics,
  Modelling
 %%SRC: 2010 CPR
 %%IML: code: YES 19940401 {LXC:DETAIL----} {UDC:(0.034)
  ,004.432,004.43.FOR} LXDATASTORAGE:///home/cpr/...
 %%IML: UDC-Object:(0.034),004.432,004.43.FOR
 %%IML: objectcomment: {PROGRAMCODE-name:
  diffam}
 %%IML: objectcomment: {PROGRAMCODE-language:
  Fortran77}
 %%IML: objectcomment: {PROGRAMCODE-compilation:
  Makefile}
 %%IML: objectcomment: {PROGRAMCODE-compiler:
  xlf}
 %%IML: objectcomment: {PROGRAMCODE-operatingsystem:
  AIX}
 %%IML: objectcomment: {PROGRAMCODE-compiler:
  g77}
 %%IML: objectcomment: {PROGRAMCODE-operatingsystem:
  Linux}
 %%IML: objectcomment: {PROGRAMCODE-compiler:
  g77}
 %%IML: objectcomment: {PROGRAMCODE-virtualsystem:
  VMWARE SuSE Linux}
 %%IML: objectcomment: {PROGRAMCODE-routine:
  ...}
 %%IML: objectcomment: {PROGRAMCODE-framework:
  ...}
 %%IML: objectcomment: {PROGRAMCODE-workflow:
  ...}
 %%IML: objectcomment: {PROGRAMCODE-usage:
  ...}
 %%IML: cite: NO  20130000 {LXK:Diffraction Amplitudes;
  Optics; Seismics;} {UDC:...} {PAGE:----..----} LXCITE:
  //Rueckemann:2013:Diffraction
```

Figure 4. Knowledge Resources, factual and conceptual complements: Natural sciences, phenomena, formalisation, and methods (excerpt).

The object carries factual knowledge and respective conceptual knowledge references. The conceptual knowledge refers to optics, diffraction, geophysics, applied geology, geological prospecting and exploration, and interpretation of results.

C. Complementary archaeology and mythology

An example object from the object collections for natural sciences and humanities of the knowledge resources' complements is shown in Figure 5.

```
Hephaistos [Archaeology, Volcanology]:
 (greek) God.
 Greek god, forger for the gods.
 Later god of fire and the forge.
 %%SRC: 1990 CPR
 compare Vulcanus
 %%IML: UDC:[902+903+904]:[25+930.85]"63"(4)(093)=14
```

Figure 5. Knowledge Resources, factual and conceptual complements: Volcanology, archaeology, and mythology (excerpt).

The object carries factual knowledge and respective conceptual knowledge references. The conceptual knowledge refers to archaeology, prehistory, prehistoric remains, artefacts, antiquities, cultural remains of historical times, religions of antiquity, minor cults and religions, history of civilization, cultural history, archaeological, prehistoric, protohistoric periods and ages, auxiliaries of place (Europe) historical sources, Greek (Hellenic) language.

An associated example object from the object collections for natural sciences and humanities of the knowledge resources' complements is shown in Figure 6.

```
Kukulkán [Archaeology]:
 (maya.) God.
 Feather Snake.
 Popol Vuh.
 %%SRC: 1990 CPR
 %%IML: UDC:[902+903+904]:[25+930.85]"63"(7)(093)=84/=88
 Syn.: Kukulcán
 s. Popol Vuh, Chichén Itzá
```

Figure 6. Knowledge Resources, factual and conceptual complements: Archaeology, mythology (excerpt).

The object carries factual knowledge and respective conceptual knowledge references. The conceptual knowledge refers to comparable context as the previous object but extends the knowledge matrix by auxiliaries of place (North and Central America) and Central and South American indigenous languages.

D. Complementary location

An associated example object from the object collections for location references of the knowledge resources' complements is shown in Figure 7.

```
L'Anse-aux-Meadows [Archaeology]:
 Viking Settlement, Newfoundland, America.
 Founded before \isodate{1000}{}{}.
 %%IML: UDC:[904]:[930.85](23)(4)(7)
 %%SRC: 1992 CPR
```

Figure 7. Knowledge Resources, factual and conceptual complements: Archaeology, physical and geographic locations (excerpt).

The object carries factual knowledge and respective conceptual knowledge references. The conceptual knowledge refers to cultural remains of historical times, history of civilization, Cultural history, North and Central America – Europe association, and above sea level context.

E. Further complementary knowledge gamut

Universal knowledge from any context is further available in order to support processes and applications besides these small examples from complementary knowledge resources. The disciplines themselves are not limited and span all knowledge, e.g., natural sciences, seismics and seismology, cartography, remote sensing, georeferences, volcanology, mineralogy, physics, chemistry, geology, mathematics,, archaeology, planetology, astrophysics, space research, biology, palaeontology, geography, religion and mythology, art, linguistics, documentation, publication.

In the context of application they also have to build facets, embracing standards and languages as well as referencing arbitrary factual knowledge, e.g., cities, countries, researchers, institutions, bibliographies.

Therefore, the comprehensive subset for geo-spatial application scenarios, which is the result of this research does have to handle these references.

F. Creation of objects and conceptual knowledge

In respect of application scenarios, e.g., object processing, the value of concordances may be reminded. Practical creation of objects has shown to be most efficient when three different categories of creation are considered:

- Manually created objects,
- Hybrid (semi-automatically) created objects, and
- Automatically created objects.

In any case creating objects is supported by universal classification, e.g., references to UDC. Therefore, that can also be applied for the creating concordances with objects.

We use a well known object for demonstration, which is referenced in the same container with all volcano objects. The listing in Figure 8 shows an instance of a simple object excerpt from an object collection. The excerpt shows keywords, content, e.g., including references, documentation, factual knowledge, and conceptual knowledge.

```
Vesuvius [Volcanology, Geology, Archaeology]:
 (lat.) Mons Vesuvius.
 (ital.) Vesuvio.
 (deutsch.) Vesuv.
 Volcano, Gulf of Naples, Italy.
 Complex volcano (compound volcano).
 Stratovolcano, large cone (Gran Cono).
 Volcano Type: Somma volcano,
 VNUM: 0101-02=,
 Summit Elevation: 1281 m.
 The volcanic activity in the region is observed by the
 Oservatorio Vesuviano. The Vesuvius area has been
 declared a national park on 1995-06-05.
 The most known antique settlements at the Vesuvius are
 Pompeji and Herculaneum.
 Syn.: Vesaevus, Vesevus, Vesbius, Vesvius
 s. volcano, super volcano, compound volcano
 s. also Pompeji, Herculaneum, seismology
 compare La Soufrière, Mt. Scenery, Soufriere
 ...
 UDC:[911.2+55]:[930.85]:[902]"63"(4+37+23+24)=12
 ...
 UCC:UDC2012:551.21
 UCC:UDC2012:551
 UCC:UDC2012:902/908
 UCC:MSC2010:86,86A17,86A60
 UCC:LCC:QE521-545
 UCC:LCC:QE1-996.5
 UCC:LCC:QC801-809
 UCC:LCC:CC1-960,CB3-482
 UCC:PACS2010:91.40.-k
 UCC:PACS2010:91.65.-n,91.
```

Figure 8. Processed instance of a simple object (excerpt) from an object collection.

Both classification and concordances, the Universal Classified Classification (UCC), were collected and created semi-automatically over a period of time. The rest of the object was created manually.

The listing in Figure 9 shows an instance of a simple container entry excerpt from a volcanological features container, in a representation, which can be input for processing workflows.

The excerpt shows a representation of conceptual knowledge for the container and various factual knowledge. The data was collected and created semi-automatically over a period of time.

```
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:UDC2012:551.21
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:UDC2012:551
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:UDC2012:551
 .2,551.23,551.24,551.26
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:UDC2012:902/908
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:MSC2010:86,86A17,86
 A60
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:LCC:QE521-545
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:LCC:QE1-996.5
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:LCC:QC801-809
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:LCC:CC1-960,CB3-482
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:PACS2010:91.40.-k
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:PACS2010:91.65.-n,91.
CONTAINER_CONCEPTUAL_KNOWLEDGE: UCC:PACS2010:91.40.Ge
 ,91.40.St,91.40.Rs,*91.45.C-,*91.45.D-,90
...
CONTAINER_OBJECT_EN_ITEM: Vesuvius
CONTAINER_OBJECT_DE_ITEM: Vesuv
CONTAINER_OBJECT_EN_PRINT: Vesuvius
CONTAINER_OBJECT_DE_PRINT: Vesuv
CONTAINER_OBJECT_EN_COUNTRY: Italy
CONTAINER_OBJECT_DE_COUNTRY: Italien
CONTAINER_OBJECT_EN_CONTINENT: Europe
CONTAINER_OBJECT_DE_CONTINENT: Europa
CONTAINER_OBJECT_XX_LATITUDE: 40.821N
CONTAINER_OBJECT_XX_LONGITUDE: 14.426E
CONTAINER_OBJECT_XX_HEIGHT_M: 1281
CONTAINER_OBJECT_EN_TYPE: Complexvolcano
CONTAINER_OBJECT_DE_TYPE: Komplex-Vulkan
CONTAINER_OBJECT_XX_VNUM: 0101-02=
```

Figure 9. Processed instance of a simple container entry (excerpt).

The conceptual knowledge of several concordances' references is exported into this representation and illustrates how to integrate conceptual knowledge implementations.

The conceptual knowledge is a matter of more detailed discussion in the next subsections and sections. The excerpts have been processed with the appropriate `lx_object_volcanology` and `lx_container_volcanology` interfaces, selecting a number of items and for the container also items in English and German including a unique formatting.

The resources' access and processing can be done in any programming language, assuming that the interfaces are implemented. For example, combining scripting, filtering, and parallel programming can provide flexible approaches.

VI. Resulting conceptual knowledge solution

Table I contains the compilation of a general comprehensive subset of resulting major conceptual knowledge reference divisions for geo-spatial application scenarios. All the conceptual knowledge reference divisions presented are referring to UDC codes, which have been made publicly available. Here, "UDC:" is the designated notation of references used with Knowledge Resources and objects in ongoing projects. The UDC illustrates the width and depth of knowledge dimensions. The full details of organisation and knowledge are available from the UDC. As far a possible, the original verbal descriptions (English for demonstration) were taken, even if the writing of terms and words may differ from the practice used for the rest of this paper. The resulting conceptual knowledge solution comprises a most comprehensive knowledge compendium of geo-spatially dominated facetted knowledge, which can be effectively and efficiently used in geo-spatial application scenarios. Besides the level of detail and arbitrary facetted knowledge, the respective conceptual knowledge reference divisions provide a focussed discipline coverage while spanning a large width and depth of knowledge reference divisions. For example, let us take an additional view on depth for UDC:004 (Computer science and technology. Computing. Data processing), UDC:51 (Mathematics), and UDC:528 (Geodesy. Surveying. Photogrammetry. Remote sensing. Cartography).

Besides the shown references, UDC:004 also comprises important subdivision context of data and structure, e.g., data handling (UDC:004.62), files (UDC:004.63), databases and their structures (UDC:004.65), and systems for numeric data (UDC:004.67). For practical references, UDC:004 can be used to also hold references to many application scenarios, e.g., algorithms for program construction, low level as well as high level and problem oriented languages, knowledge representation, artificial intelligence application systems, intelligent knowledge-based systems. For practical references with mathematical, geometrical, and topological context, UDC:51 can be used to also hold references to fundamental and general considerations of mathematics, number theory, algebra, geometry, topology, analysis, combinatorial analysis, graph theory, probability, mathematical statistics, computational mathematics, numerical analysis, mathematical cybernetics, operational research as well as mathematical theories and methods. For practical references with geoscience and spatial disciplines, UDC:528 can be used to also hold references to a much deeper discipline based knowledge, e.g., fundamentals derived from potential theory, level surfaces, geoids, geometric/static methods, use of longitudinal and latitudinal measurements, gravity measurement, astro-geodetic determination of position, geographical coordinates, topographic surveying, engineering surveys, special fields of surveying, applications of photogrammetry, fundamental and physical principles, data processing, and interpretation.

The result of conceptual knowledge reference divisions based on the methodology of superordinate knowledge is complementary to geo-spatial topologies and geo-referencing. It can be used complementary with any geoscientific and geo-spatial knowledge in any context.

The result can provide solutions wherever conceptual knowledge references are involved. The methodologies and implementations make sure that powerful sets of unique attributes and features are available. The number of possible use cases is practically unlimited. The case studies showed that a wide range of application scenarios can benefit from the principles of superordinate knowledge and considering conceptual knowledge as complementary means for consistently documenting and handling knowledge. The passages in the following section refer to discussions and details for an excerpt of successful implementations.

VII. Evaluation from implementation cases

Many years of research and practical solution developments contributed to creating a comprehensive subset of conceptual knowledge, which is the fundament deployed for general practical solutions, e.g., with geo-spatial applications and with geo-data knowledge mining and processing.

TABLE I. COMPREHENSIVE SUBSET OF RESULTING CONCEPTUAL KNOWLEDGE REFERENCE DIVISIONS FOR GEO-SPATIAL APPLICATION SCENARIOS, PRACTICALLY USED MAIN CLASSIFICATION REFERENCES, UNIVERSAL DECIMAL CLASSIFICATION SAMPLES (UDC, ENGLISH; UDCC [17]; CC [18]).

CONCEPTUAL KNOWLEDGE REFERENCES FOR GEO-SPATIAL SCENARIOS			
Code / Sign Ref.	*Verbal Description (EN)*	*Code / Sign Ref.*	*Verbal Description (EN)*
	Common Auxiliary Signs		
+	Coordination. Addition (plus sign).	[]	Subgrouping (square brackets).
/	Consecutive extension (oblique stroke sign).	*	Introduces non-UDC notation (asterisk).
:	Simple relation (colon sign).	A/Z	Direct alphabetical specification.
::	Order-fixing (double colon sign).	,	[Reference listing, itemisation]
	Auxiliary Tables		
UDC:=...	Common auxiliaries of language.	UDC:(=...)	Common auxiliaries of human ancestry, ethnic grouping and nationality.
UDC:(0...)	Common auxiliaries of form.		
UDC:(1/9)	Common auxiliaries of place.	UDC:-0...	Common auxiliaries of general characteristics: Properties, Materials, Relations/Processes and Persons.
UDC:"..."	Common auxiliaries of time.		
	Place and Space		
UDC:(1/9)	Common auxiliaries of place.	UDC:(20)	Ecosphere
UDC:(1)	Place and space in general. Localization. Orientation	UDC:(21)	Surface of the Earth in general. Land areas in particular. Natural zones and regions
UDC:(100)	Universal as to place. International. All countries in general		
UDC:(1-0/-9)	Special auxiliary subdivision for boundaries and spatial forms of various kinds	UDC:(23)	Above sea level. Surface relief. Above ground generally. Mountains
UDC:(1-0)	Zones		
UDC:(1-1)	Orientation. Points of the compass. Relative position	UDC:(24)	Below sea level. Underground. Subterranean
UDC:(1-2)	Lowest administrative units. Localities	UDC:(25)	Natural flat ground (at, above or below sea level). The ground in its natural condition, cultivated or inhabited
UDC:(1-5)	Dependent or semi-dependent territories		
UDC:(1-6)	States or groupings of states from various points of view	UDC:(26)	Oceans, seas and interconnections
UDC:(1-7)	Places and areas according to privacy, publicness and other special features	UDC:(28)	Inland waters
		UDC:(29)	The world according to physiographic features
UDC:(1-8)	Location. Source. Transit. Destination	UDC:(3/9)	Individual places of the ancient and modern world
UDC:(1-9)	Regionalization according to specialized points of view	UDC:(3)	Places of the ancient and mediaeval world
UDC:(2)	Physiographic designation	UDC:(4/9)	Countries and places of the modern world
	Main Tables		
UDC:0	Science and Knowledge. Organization. Computer Science. Information. Documentation. Librarianship. Institutions. Publications	UDC:5	Mathematics. Natural Sciences
		UDC:6	Applied Sciences. Medicine, Technology
UDC:1	Philosophy. Psychology	UDC:7	The Arts. Entertainment. Sport
UDC:2	Religion. Theology	UDC:8	Linguistics. Literature
UDC:3	Social Sciences	UDC:9	Geography. Biography. History
	Science, Knowledge, Organisation		
UDC:001	Science and knowledge in general. Organization of intellectual work	UDC:007	Activity and organizing. Communication and control theory generally (cybernetics). 'Human engineering'
UDC:002	Documentation. Books. Writings. Authorship		
UDC:003	Writing systems and scripts	UDC:01	Bibliography and bibliographies. Catalogues
UDC:004	Computer science and technology. Computing. Data processing	UDC:02	Librarianship
UDC:004.4	Software	UDC:030	General reference works (as subject)
UDC:004.6	Computer data	UDC:050	Serial publications, periodicals (as subject)
UDC:004.7	Computer communication. Computer networks	UDC:06	Organizations of a general nature
UDC:004.8	Artificial intelligence	UDC:061	Organizations and other types of cooperation
UDC:005	Management	UDC:069	Museums. Permanent exhibitions
UDC:005.94	Knowledge management	UDC:070	Newspapers (as subject). The Press. Journalism
UDC:006	Standardization of products, operations, weights, measures and time	UDC:08	Polygraphies. Collective works
UDC:008	Civilization. Culture. Progress	UDC:09	Manuscripts. Rare and remarkable works
	Geo-spatial Focus Divisions From Main Tables		
UDC:51	Mathematics	UDC:550.3	Geophysics
UDC:528	Geodesy. Surveying. Photogrammetry. Remote sensing. Cartography	UDC:550.7	Geobiology. Geological actions of organisms
UDC:528.2	Figure of the Earth. Earth measurement. Mathematical geodesy. Physical geodesy. Astronomical geodesy	UDC:550.8	Applied geology and geophysics. Geological prospecting and exploration. Interpretation of results
UDC:528.3	Geodetic surveying	UDC:551	General geology. Meteorology. Climatology. Historical geology. Stratigraphy. Palaeogeography
UDC:528.4	Field surveying. Land surveying. Cadastral survey. Topography. Engineering survey. Special fields of surveying	UDC:551.8	Palaeogeography
UDC:528.7	Photogrammetry: aerial, terrestrial	UDC:778	Special applications and techniques of photography
UDC:528.8	Remote sensing	UDC:91	Geography. Exploration of the Earth and of individual countries. Travel. Regional geography (systematic geography). Theoretical geography
UDC:528.9	Cartography. Mapping (textual documents)		
UDC:528.94	Thematic cartography. Topical cartography		
UDC:53	Physics	UDC:912	Nonliterary, nontextual representations of a region
UDC:55	Earth Sciences. Geological sciences	UDC:913	Regional geography

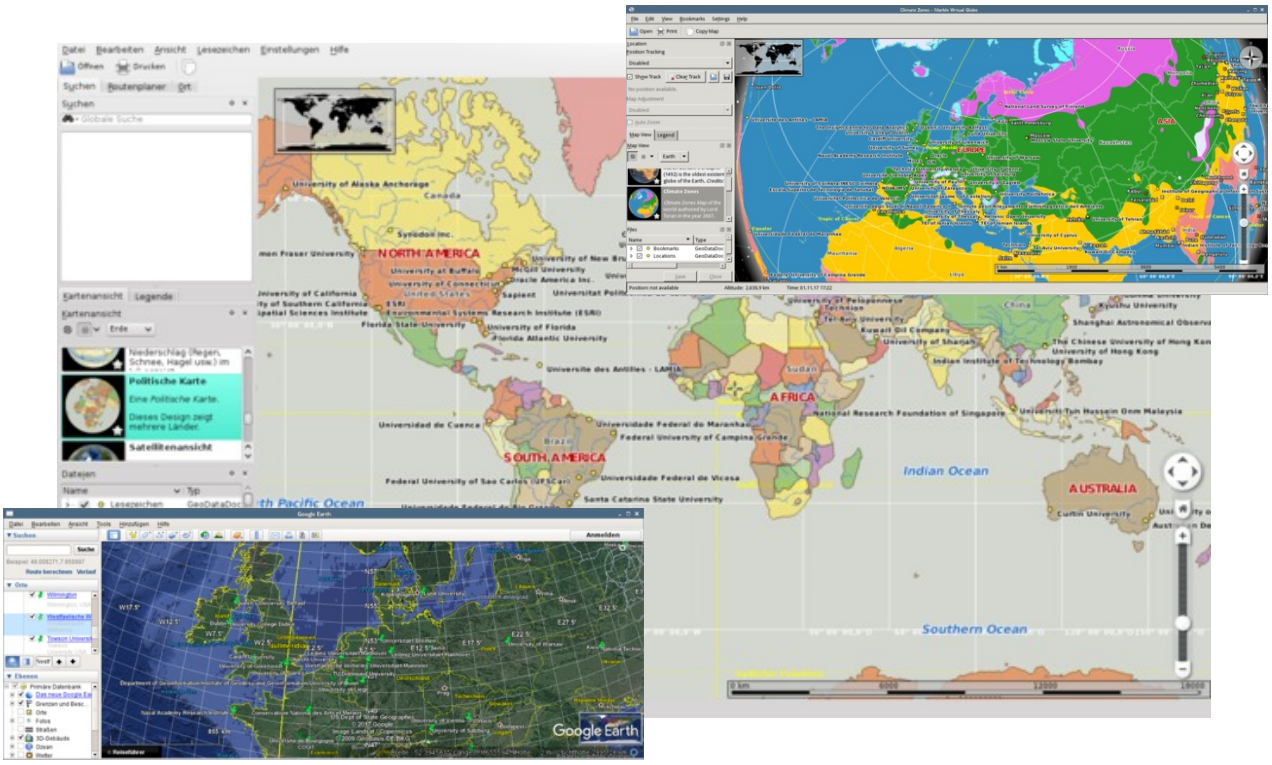


Figure 10. Collage of different implementation cases based on the resulting conceptual knowledge (Table I) from this research: Knowledge mapping, integration, mining; the samples illustrate context creation and dynamical visualisation. For technical details please see the references for the case studies given in the text.

The conceptual knowledge framework employed here, especially UDC, has passed the test of time and is so mature and used in so many scenarios that the ongoing knowledge development itself is iterating with its application.

The solution showed to fulfill all the major significant criteria, especially:

- The conceptual knowledge references (e.g., UDC) and components of the Knowledge Resources fully support for universal knowledge.
- These references and components of the Knowledge Resources fully support multi-disciplinary solutions, integrating facetted universal knowledge. All the components are multi-lingual. The conceptual knowledge references currently support about fifty languages.
- All the components, including the conceptual knowledge references allow sustainable, long-term usable edition framework, which is a base for consistent, extendable solutions on all components.

The solution allows all the consequent advanced features, which are directly linked to knowledge and documentation.

The previously unpublished results of practical conceptual knowledge are first presented with the research cited in this paper (Table I). The following case studies are based on these results and present small but illustrative excerpts (Figure 10) in form of a cross-section of conducted research and development, of Knowledge Resources, algorithms, intelligent workflows, and implementations.

Here, for example, a knowledge mining process employing knowledge objects based on the referred conceptual knowledge can use all the width and depth of knowledge behind the comprehensive subset to automatically or semi-automatically create new context and visualisation for a data set containing non-georeferenced text entities (affiliations in floating text), e.g., geographical, political, and climate zone context.

Besides the knowledge fundament and framework being focus of this research paper, the references in the next passages contain further details for the practical case studies, the implemented methods and the technologies, which were used for the different case studies.

- *Knowledge integration* allows to create new views and insights by computing Spatial Cogwheel modules [32].
- *Knowledge mining:* Creating Knowledge Resources and employing classification and concordances can provide a base for advanced knowledge discovery and computational solutions [16]. The integration of Knowledge Resources and advanced association processing can be beneficial in many disciplines as it provides multi-disciplinary and multi-lingual support [33].
  Methods like the Content Factor can be used for advanced knowledge processing [34]. The integration of appropriate methods can be used for further advancing the Knowledge Resources, as well as the mining processes [35].

- *The methodology of knowledge mapping* allows to create flexible methods in order to handle spatial representations and knowledge mining by creating a multi-dimensional context for arbitrary objects and entities [14].
- *Dynamical visualisation:* The methodology can be used for enabling knowledge based methods for computation and computational and dynamical visualisation [36].
- *Association and phonetic features:* The methodology supports phonetic association and mining methods [37].
- *Verbal description:* The employment of implemented methods can be supported and make use of multi-lingual verbal descriptions and concordances [38] as the conceptual knowledge is consistently available in 50 languages, providing millions of basic conceptual knowledge references.

VIII. Conclusion

This paper discussed the practice of formalised conceptual knowledge complements created with multi-disciplinary collections, containers, and referenced resources and presented different types of multi-disciplinary Knowledge Resources.

This paper presented a complements of formalised conceptual knowledge and a representative compilation of components from successful complex implementation scenarios. All such components are under creation and development for several decades, which have shown that the creation and development of multi-disciplinary Knowledge Resources is an essential long-term value.

With this research, a comprehensive subset of conceptual knowledge reference divisions was created, further developed, and finally compiled from the practical application case studies, which have been conducted and further developed over the last decades. This research achieved to create a comprehensive tool base of conceptual knowledge in geo-spatial application scenarios for all kinds of multi-disciplinary knowledge context in any media. The implemented superordinate knowledge based solution fulfills all the required criteria as was presented and discussed in this paper.

The result was employed to successfully implement a wide range of different geo-spatial cases.

Based on the presented research and practiced during extensive creation and development of complementary knowledge resources, a comprehensive subset of references to conceptual knowledge, allowing geo-spatially dominated facetted knowledge, was created, further developed, and finally compiled from the application case studies, which have been conducted over the last three decades. Knowledge based fundaments, e.g., those built on UDC, showed to have a very high impact on knowledge creation and mining in theory and practice, not only for spatial knowledge.

The knowledge approach proved to be a fundamental "enabler" and contributed significantly to many solutions. Covering a coherent system of knowledge provides a holistic and consistent environment for any scenario, which is supported by excellent features for facetted knowledge. The referenced conceptual knowledge itself is consistent due to its development and publication via editions. Implementations support fully multi-disciplinary context and multi-lingual instances for many languages. Solutions are extensible to integrate and fit special purposes. The methodology is data-centric and scalable for width and depth of knowledge as well as for infrastructure requirements. All the cases so far implementing the presented solution provided seamless integration with common geo-spatial practices and showed excellent sustainability, knowledge coverage, long-term characteristics, and scalability. In review of these results, all major institutions, e.g., libraries focussing on information science and research data management, are using and developing conceptual knowledge with their core tasks, which opens up a wide range of excellent knowledge sources, which can be considered high value resources. Moreover, such Knowledge Resources are complementary, independent of the fact that they can incorporate different methods and approaches, e.g., thesauri, semantic frameworks, ontologies, and phonetic interfaces for the content they handle.

Future research on theory and practice will concentrate on further developing the spectrum of references and creating knowledge reference based solutions for scenarios and disciplines. In addition, the creation and further development of knowledge resources is a multi-disciplinary long-term task.

Acknowledgements

We are grateful to the "Knowledge in Motion" (KiM) long-term project, Unabhängiges Deutsches Institut für Multidisziplinäre Forschung (DIMF), for partially funding this research, implementation, case studies, and publication under grants D2016F5P04648 and D2018F6P04938 and to its senior scientific members and members of the permanent commission of the science council, and the board of trustees, especially to Dr. Friedrich Hülsmann, Gottfried Wilhelm Leibniz Bibliothek (GWLB) Hannover, to Dipl.-Biol. Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, to Dipl.-Ing. Martin Hofmeister, Hannover, and to Olaf Lau, Hannover, Germany, for collaboration, practical multi-disciplinary case studies, and the analysis of advanced concepts. We are grateful to Dipl.-Ing. Hans-Günther Müller, Cray, Germany, for his excellent contributions and assistance providing practical private cloud and storage solutions. We are grateful to all national and international partners in the Geo Exploration and Information cooperations for their constructive and trans-disciplinary support. We are grateful to the Science and High Performance Supercomputing Centre (SHPSC) for long-term support. / DIMF-PIID-DF98_007.

References

[1] C.-P. Rückemann, "Superordinate Knowledge Based Comprehensive Subset of Conceptual Knowledge for Practical Geo-spatial Application Scenarios," in Proceedings of The Eleventh International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2019), February 24 – 28, 2019, Athens, Greece. XPS Press, Wilmington, Delaware, USA, 2019, pp. 52–58, ISSN: 2308-393X, ISBN: 978-1-61208-687-3, URL: http://www.thinkmind.org/index.php?view=article&articleid=geoprocessing_2019_3_30_30039 [accessed: 2019-07-07].

[2] C.-P. Rückemann, Sustainable Knowledge and Resources Management for Environmental Information and Computation. Business Expert Press, Manhattan, New York, USA, Mar. 2018, Chapter 3, pp. 45–88, in: Huong Ha (ed.), Climate Change Management: Special Topics in the Context of Asia, ISBN: 978-1-94784-327-1 (paperback), ISBN: 978-1-94784-328-8 (e-book), in: Robert Sroufe (ed.), Business Expert Press Environmental and Social Sustainability for Business Advantage Collection, ISSN: 2327-333X (collection, print).

[3] "ISO 14000 - Environmental management," 2019, URL: http://www.iso.org/iso/iso14000 [accessed: 2019-05-02].

[4] C.-P. Rückemann, "Superordinate Knowledge Based Comprehensive Subset of Conceptual Knowledge for Practical Mathematical-Computational Scenarios," in The Ninth Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of The 17th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 23–28, 2019, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), no. 1. AIP Press, American Institute of Physics, Melville, New York, USA, Oct. 2020, ISSN: 0094-243X, (to appear).

[5] Aristotle, Nicomachean Ethics, Volume 1, 2009, Project Gutenberg, eBook, EBook-No.: 28626, Release Date: April 27, 2009, Digitised Version of the Original Publication, Produced by Sophia Canoni, Book provided by Iason Konstantinidis, Translator: Kyriakos Zambas, URL: http://www.gutenberg.org/ebooks/12699 [accessed: 2019-07-07].

[6] Aristotle, The Ethics of Aristotle, 2005, Project Gutenberg, eBook, EBook-No.: 8438, Rel. Date: Jul., 2005, Digit. Vers. of the Orig. Publ., Produced by Ted Garvin, David Widger, and the DP Team, Edition 10, URL: http://www.gutenberg.org/ebooks/8438 [accessed: 2019-07-07].

[7] L. W. Anderson and D. R. Krathwohl, Eds., A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon, Boston, MA (Pearson Education Group), USA, 2001, ISBN: 978-0801319037.

[8] C.-P. Rückemann, F. Hülsmann, B. Gersbeck-Schierholz, P. Skurowski, and M. Staniszewski, Knowledge and Computing. Post-Summit Results, Delegates' Summit: Best Practice and Definitions of Knowledge and Computing, Sept. 23, 2015, The Fifth Symp. on Adv. Comp. and Inf. in Natural and Applied Sciences (SACINAS), The 13th Int. Conf. of Num. Analysis and Appl. Math. (ICNAAM), Sept. 23–29, 2015, Rhodes, Greece, 2015, DOI: 10.15488/3409, URL: https://doi.org/10.15488/3409 [accessed: 2019-07-07].

[9] Plato, Phaedo, 2008, (Written 360 B.C.E.), Translated by Benjamin Jowett, Provided by The Internet Classics Archive, URL: http://classics.mit.edu/Plato/phaedo.html [accessed: 2019-07-07].

[10] T. Gooley, How to Read Nature: Awaken Your Senses to the Outdoors You've Never Noticed. New York, N.Y.: Experiment, 2017, ISBN: 978-1-61519-429-2.

[11] J. Burroughs, Leaf and Tendril, 1908, Ch. 1, The Art of Seeing Things.

[12] C.-P. Rückemann, "Principles of Superordinate Knowledge: Separation of Methodology, Implementation, and Realisation," in The Eighth Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of The 16th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 13–18, 2018, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), vol. 2116. AIP Press, American Institute of Physics, Melville, New York, USA, 2019, ISSN: 0094-243X, (to appear).

[13] C.-P. Rückemann and F. Hülsmann, "Significant Differences: Methodologies and Applications," "Significant Differences: Methodologies and Applications", KiMrise, Knowledge in Motion Meeting, November 27, 2017, Knowledge in Motion, Hannover, Germany, 2017.

[14] C.-P. Rückemann, "Methodology of Knowledge Mapping for Arbitrary Objects and Entities: Knowledge Mining and Spatial Representations – Objects in Multi-dimensional Context," in Proceedings of The Tenth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2018), March 25–29, 2018, Rome, Italy. XPS Press, Wilmington, Delaware, USA, 2018, pp. 40–45, ISSN: 2308-393X, ISBN: 978-1-61208-617-0, URL: http://www.thinkmind.org/index.php?view=article&articleid=geoprocessing_2018_3_20_30078 [accessed: 2019-07-07].

[15] Organisation for Economic Co-operation and Development (OECD), "OECD Principles and Guidelines for Access to Research Data from Public Funding," 2007, URL: https://www.oecd.org/sti/sci-tech/38500813.pdf [accessed: 2019-07-07].

[16] C.-P. Rückemann, "Advanced Knowledge Discovery and Computing based on Knowledge Resources, Concordances, and Classification," International Journal on Advances in Intelligent Systems, vol. 9, no. 1&2, 2016, pp. 27–40, ISSN: 1942-2679, URL: http://www.thinkmind.org/index.php?view=article&articleid=intsys_v9_n12_2016_3/ [accessed: 2019-07-07].

[17] "Multilingual Universal Decimal Classification Summary," 2012, UDC Consortium, 2012, Web resource, v. 1.1. The Hague: UDC Consortium (UDCC Publication No. 088), URL: http://www.udcc.org/udcsummary/php/index.php [accessed: 2019-07-07].

[18] "Creative Commons Attribution Share Alike 3.0 license," 2012, URL: http://creativecommons.org/licenses/by-sa/3.0/ [accessed: 2019-07-07], (first release 2009, subsequent update 2012).

[19] F. Hülsmann and C.-P. Rückemann, "Symbolism and Meaning and Their Significance in Information Science," KiM Summit, February 20, 2018, Knowledge in Motion, Hannover, Germany, 2019.

[20] B. Gersbeck-Schierholz, "Art and Symbolism: The Sistine Chapel," KiM On-site Summit, Knowledge in Motion, March 27, 2018, On-site Summit Meeting, "Unabhängiges Deutsches Institut für Multidisziplinäre Forschung (DIMF)", Rome, Italy, 2018.

[21] B. Gersbeck-Schierholz and C.-P. Rückemann, "To See and Not to See," KiM On-site Summit, Knowledge in Motion, July 26, 2018, On-site Summit Meeting, "Unabhängiges Deutsches Institut für Multidisziplinäre Forschung (DIMF)", Museu Nacional d'Art de Catalunya, Barcelona, Spain, 2018.

[22] "UDC, Common Auxiliary Signs," 2019, Universal Decimal Classification (UDC), URL: https://udcdata.info/078885 [accessed: 2019-07-07].

[23] "UDC Summary Linked Data, Main Tables," 2019, Universal Decimal Classification (UDC), URL: https://udcdata.info/078887 [accessed: 2019-07-07].

[24] "UDC 0: Science and knowledge. Organization. Computer science. Information. Documentation. Librarianship. Institution. Publications," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/13358 [accessed: 2019-07-07].

[25] "UDC (1/9): Common auxiliaries of place," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/001951 [accessed: 2019-07-07].

[26] "UDC 528: Geodesy. Surveying. Photogrammetry. Remote sensing. Cartography," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/027504 [accessed: 2019-07-07].

[27] "UDC 910: General questions. Geography as a science. Exploration. Travel," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/068129 [accessed: 2019-07-07].

[28] "UDC 912: Nonliterary, nontextual representations of a region," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/068183 [accessed: 2019-07-07].

[29] "UDC 913: Regional geography," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/068186 [accessed: 2019-07-07].

[30] "UDC 94: General history," 2019, Universal Decimal Classification (UDC), URL: http://udcdata.info/068284 [accessed: 2019-07-07].

[31] N. T. Nikolitsis, "Archäologische Unterwasser-Expedition bei Rhodos, (English: Archaeological Underwater-Expedition at Rhodes)," Antike Welt, Zeitschrift für Archäologie und Kulturgeschichte, (English: Antique World, Magazine for Archaeology and Cultural History), 1981, 12. Jg., Heft 1, (English: 12th Year, Issue 1), pp. 45–58.

[32] C.-P. Rückemann, "Creating New Views and Insights by Computing Spatial Cogwheel Modules for Knowledge Integration," Int. Journ. on Adv. in Intell. Syst., vol. 10, no. 3&4, 2017, pp. 314–326, ISSN: 1942-2679, URL: http://www.thinkmind.org/index.php?view=article&articleid=intsys_v10_n34_2017_13/ [accessed: 2019-07-07].

[33] C.-P. Rückemann, "Integration of Knowledge Resources and Advanced Association Processing for Geosciences and Archaeology," Int. Jour. on Adv. in Systems and Measurements, vol. 9, no. 3&4, 2016, pp. 485–495, ISSN: 1942-261x, URL: http://www.thinkmind.org/index.php?view=article&articleid=sysmea_v9_n34_2016_22/ [accessed: 2019-07-07].

[34] C.-P. Rückemann, "Knowledge Processing and Advanced Application Scenarios With the Content Factor Method," International Journal on Advances in Intelligent Systems, vol. 9, no. 3&4, 2016, pp. 485–495, ISSN: 1942-2679, URL: http://www.thinkmind.org/index.php?view=article&articleid=intsys_v9_n34_2016_22/ [accessed: 2019-07-07].

[35] C.-P. Rückemann, "Progressive Advancement of Knowledge Resources and Mining: Integrating Content Factor and Comparative Analysis Methods for Dynamical Classification and Concordances," Int. Journal on Adv. in Systems and Measurements, vol. 11, no. 1&2, 2018, ISSN: 1942-261x, URL: http://www.thinkmind.org/index.php?view=article&articleid=sysmea_v11_n12_2018_5/ [accessed: 2019-07-07].

[36] C.-P. Rückemann, "Creating Knowledge-based Dynamical Visualisation and Computation," in Proc. of The Seventh International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2015), February 22–27, 2015, Lisbon, Portugal. XPS Press, 2015, pp. 56–62, ISSN: 2308-393X, ISBN: 978-1-61208-383-4, URL: http://www.thinkmind.org/index.php?view=article&articleid=geoprocessing_2015_3_40_30063 [accessed: 2019-07-07].

[37] C.-P. Rückemann, "Archaeological and Geoscientific Objects used with Integrated Systems and Scientific Supercomputing Resources," Int. Jour. on Adv. in Systems and Measurements, vol. 6, no. 1&2, 2013, pp. 200–213, ISSN: 1942-261x, URL: http://www.thinkmind.org/download.php?articleid=sysmea_v6_n12_2013_15 [accessed: 2019-07-07].

[38] C.-P. Rückemann, "Methodology Enabling Knowledge Mining Computation Based on Conceptual Knowledge and Verbal Description," in The Seventh Symposium on Advanced Computation and Information in Natural and Applied Sciences (SACINAS), Proc. of The 15th Int. Conf. of Num. Analysis and Appl. Math. (ICNAAM), Sept. 25–30, 2017, Thessaloniki, Greece, AIP Conference Proceedings, Vol. 1728, no. 1. AIP Press, Melville, New York, USA, Jul. 2018, ISBN: 978-0-7354-1690-1 (Book), ISSN: 0094-243X, DOI: 10.1063/1.5043723.

Data Quality Challenges in Weather Sensor Data, Including Identification of Mis-located Sites

Douglas E. Galarus
Computer Science Department
Utah State University
Logan, UT 84322-4205, United States
douglas.galarus@usu.edu

Rafal A. Angryk
Department of Computer Science
Georgia State University
Atlanta, GA 30302, United States
angryk@cs.gsu.edu

Abstract—**There are many challenges in developing and evaluating methods including: real-world cost and infeasibility of verifying ground truth, non-isotropic covariance, near-real-time operation, challenges with time, bad data, bad metadata, and other quality factors. In this paper, we demonstrate the challenges of evaluating spatio-temporal data quality methods for weather sensor data via a method we developed and other popular, interpolation-based methods to conduct model-based outlier detection. We demonstrate that a multi-faceted approach is necessary to counteract the impact of outliers. We demonstrate the challenges of evaluation in the presence of incorrect labels of good and bad data. We also investigate, in depth, the challenge of identifying mis-located sites.**

Keywords-Data Quality; Spatial-Temporal Data; Quality Control; Outlier; Inlier; Bad Data; Ground Truth; Bad Metadata

I. INTRODUCTION

In our research, we address near-real-time determination of outliers and anomalies in spatiotemporal weather sensor data, and the implications of quality assessment on computation from the perspective of the data aggregator. This paper extends the work presented in the conference paper at GeoProcessing 2019 [1]. Data might not reflect the conditions they measure for a variety of reasons. The challenges go beyond identifying individual outlying observations. A sensor might become "stuck" and produce the same output over an extended period. A sensor's output may conform to other nearby observations and fall within an acceptable range of values, but not reflect actual conditions. A sensor may drift, reporting values further from ground truth over time. A sensor may report correct values, but the associated clock may be incorrect, resulting in bad timestamps. An incorrect location may be associated with a site. In fact, many sites may be mis-located. These and related problems cause challenges that are far more complex than simple outlier detection.

Sensor-level quality control processes often utilize domain-specific, rule-based systems or general outlier detection techniques to flag "bad" values. NOAA's Meteorological Assimilation Data Ingest System (MADIS) [2] applies the range [-60° F, 130° F] to check for air temperature observations [3] while the University of Utah's MesoWest [4] uses the range [-75° F, 135° F] [5] for validity checks. These ranges are intended to represent the possible air temperature values in real world conditions, at least within the coverage area of the provider. If an observation falls outside the range, then the provider flags that observation as having failed the range test and the observation will, for all practical purposes, be considered "bad". Range tests are not perfect. The record high United States temperature would fail MADIS's range test, although it would pass MesoWest's test. Both MADIS and MesoWest further employ a suite of tests that go beyond their simple range tests. "Buddy" tests compare an observation to neighboring observations. MADIS uses Optimal Interpolation in conjunction with cross-validation to measure the conformity of an observation to its neighbors [3]. MesoWest estimates observations using multivariate linear regression [6]. A real observation is compared to the estimate, and if the deviation is high, then the real observation is flagged as questionable.

These approaches are flawed in that they do not account for bad metadata, such as incorrect timestamps or incorrect locations. They do not account for chronically bad sites which produce bad data including data that may sometimes appear correct. Of even greater concern, they may not do a good job in assessing accuracy and may be incorrectly labeling bad data as good and good data as bad.

The consequences of ignoring data quality are great. How can we trust our applications and models if the inputs are bad? In turn, how can we better assess data for quality so that we can be confident in its use?

In this paper, we present evaluation results for our previously published method including evaluation with several data sets. These results are significant in that they demonstrate the challenges of evaluation of methods for data quality assessment of spatio-temporal weather sensor data. We also investigate in depth the problem of identifying mis-located sites. The rest of this paper is organized as follows: Section II presents relevant literature, Section III identifies general challenges, Section IV defines our approach, Section V documents evaluation results, Section VI presents new investigation of mis-located sites, and Section VII gives our conclusions.

II. LITERATURE REVIEW

The data mining process includes data preprocessing and cleaning as critical components. Outlier analysis, is addressed within these headings by Han et al. [7], and the impact of outliers is covered by Nisbet et al. [8]. Robust re-

gression techniques are employed in data mining to overcome outliers and low quality data in the process of data cleaning by Witten et al. [9]. The handling of errors and missing values is presented by Steinbach and Kumar [10], along with quality attributes, such as accuracy and precision, as well as the adverse impact that outliers can have on clustering algorithms. Such examples demonstrate the chicken-egg nature of the problem in which a method used to identify outliers is adversely impacted by outliers.

Aggarwal [11] presents a number of useful, general observations: Correlation across time series can help to identify outliers, using one or multiple series to predict another. Deviations between predicted and actual values can then be used to identify outliers. When used on temporal snapshots of data, spatial methods can fall short because they do not address the time component. Decoupling the spatial and temporal aspects can be suboptimal. Neighborhoods can be used to make predictions, yet it is a challenge to combine spatial and temporal dimensions in a meaningful way. Domain-specific methods can be used to filter noise, but such filtering can mask anomalies in the data.

Shekhar et al. [12] present a unified approach for detecting spatial outliers and a general definition for spatial outliers, but they do not address the spatio-temporal situation. Klein et al. [13]–[17] present work on transfer and management challenges related to the inclusion of quality control information in data streams and develop optimal, quality-based load-shedding for data streams in. A missing component is the spatial aspect.

The weather and road-weather communities employ detailed accuracy checks for individual observations. The Oklahoma Mesonet uses the Barnes Spatial Test [18], a variation of Inverse Distance Weighting (IDW) (see Shepard [19]). MesoWest [4] uses multivariate linear regression to assess data quality for air temperature, as described by Splitt and Horel in [20] and [21]. MADIS [2] implements multi-level, rule-based quality control checks including a level-3 neighbor check using Optimal Interpolation / kriging [3][22][23]. These approaches (IDW, Linear Regression, kriging) can be used to check individual observations for deviation from predicted and flag individual observations as erroneous or questionable if the deviation is *large*. But if interpolated values are erroneous, then the quality assessment will be bad too. If metadata, such as location or timestamps associated with a site, is erroneous, then the quality control assessment may be bad because of comparison with the wrong data from the wrong sites. None of these approaches identify incorrect location metadata. One provider, Mesowest, attempts to identify bad timestamps, yet their approach only identifies one of the most obvious timestamp-related problem – timestamps that cannot possibly be correct because they occur in the future relative to collection time. Our own experience in this domain has been that sites are often mis-located. We found in one instance that multiples sites were mis-located with different locations across four systems from which we were extracting data.

Many spatial approaches use interpolation for quality assessment, so it is useful to examine work that compares and enhances traditional interpolation methods. Zimmerman et al. [24] use artificial surfaces and sampling techniques, as well as noise level and strength of correlation, to compare Ordinary kriging (OK) and Universal kriging (kriging with a trend) (UK) and IDW. They found that the kriging methods outperformed IDW across all variations they examined. Lu and Wong [25] found instances in which kriging performed worse than their modified version of IDW, where they vary the exponent depending on the neighborhood. They indicate that kriging would be favored in situations for which a variogram accurately reflects the spatial structure. Mueller et al. [26] show similar results, saying that IDW is a better choice than OK in the absence of semi-variograms to indicate spatial structure.

In prior work, we proposed a modification of IDW that used a data-based distance rather than geographic distance to assess observation quality [27][28]. That work focused on the use of robust methods to associate sites for assessment of individual observations. In [29][30][31], we extended the mappings to better account for spatio-temporal variation and observation time differences when assessing observations. In [32] and [33], we developed quality measures that extended beyond sites, to help evaluate overall spatial and temporal coverage of a region.

IDW is widely applied, including applications which involve outlier detection and mitigation. Xie et al. [34] applied it to surface reconstruction, in which they detect outliers using distance from fitted surfaces. Others extend the method in different ways including added dimensions, particularly time. Li et al. extend IDW in [35] to include the time dimension in their application involving estimated exposure to fine particulate matter. Grieser warns of problems with arbitrarily large weights when sites are near in analyzing monthly rain gauge observations [36], and mitigates the problem in a manner that Shepard originally used by defining a neighborhood for which included points are averaged with identical weights in place of the large, inverse distance weights.

Kriging and Optimal Interpolation were developed separately and simultaneously as spatial best linear unbiased predictors (blups) that are for practical purposes equivalent. L. S. Gandin, a meteorologist, developed and published optimal interpolation in the Soviet Union in 1963. Georges Matheron, a French geologist and mathematician, developed and published kriging in 1962, named for a South African mining engineer, Danie Krige, who partially developed the technique in 1951 and later in 1962. For further information, refer to Cressie [37].

Kriging is easily impacted by multiple data quality dimensions and its applicability is hindered unless data quality issues in the inputs are addressed. Kriging will downweight observations that are clustered in direction, as indicated by Wackernagel et al. [38]. This may be beneficial. However, a near observation can also shadow far observations in the same direction, causing them to have small or

even negative weights. This is problematic in the case that the near observation is bad.

Kriging is typically used to interpolate values at locations for which measurements are unknown using observations from known locations. As such, covariance is typically estimated. This estimate usually takes the form of a function of distance alone and is determined by the data set. A principal critique of kriging is that while it does produce optimal results when the covariance structure is known, the motivation for using kriging is questionable when the covariance structure must be estimated. Handcock and Stein [39] make such an argument. Another critique is that kriging will yield a model that matches data input to the model, giving the (false) impression that the model is perfect, as stated by Hunter et al. [40].

Unfortunately, none of these approaches alone directly addresses outlier and anomaly detection for spatio-temporal data in a robust and comprehensive manner that meets our needs. None identify bad sites and metadata in a comprehensive manner. Even so, the data quality attributes presented are of some benefit and the methods used by the weather data providers appear to be state of the art for assessment of accuracy.

III. CHALLENGES

Our research involves (fixed) site-based, spatio-temporal sensor big data, acquired and evaluated for data quality with real-time potential. There are many computational challenges associated with our problem. We focus subsequent evaluation on scalability and accuracy.

Scalability. Our data sets include thousands of sites, with potential to expand to tens of thousands of sites. Sites have varying reporting frequencies ranging from every minute to hourly or longer. These sites collectively generate millions of observations daily. We desire to run our algorithms in near real-time, and scalability is key to achieving this goal.

Accuracy. The underlying data has many data quality challenges. Accurately modeling the data is challenging, because the modeled data will inherently include errors. We desire robust, accurate models that can be used to assess the quality of individual observations.

There are many indirect issues causing challenges that must be overcome. These all influence or are influenced by computation in one way or another.

Real-World Cost and Infeasibility of Verifying Ground Truth. Agencies cannot verify ground truth on a regular basis across hundreds or thousands of sites. Human-required resolution processes can be focused if problems are identified automatically. Third-party data aggregators have no control over original data quality. Assessment of quality is essential for use.

Non-Isotropic Covariance. Distance cannot be treated equally in all dimensions nor in all directions. There are differences between the time dimension and spatial dimensions. Elevation, proximity to the ocean, terrain, microclimates, prevailing weather patterns, the diurnal effect, seasonal change, etc. also cause differences in covariance.

Near-Real-Time Operation. We intend for our processes to run in near-real-time when observations are acquired. We store and use only the most recent observations for near real-time presentation and comparison. We do not intend to store third-party historical data on our production systems. This does not preclude the potential for offline preprocessing and analysis that makes use of historical data. Even if providers apply their own quality control measures, near-real-time operation may require us to use observations that have not been fully quality-checked.

Further Challenges with Time. Sites report observations at discrete times resulting in granularity and non-uniformity. Observation frequencies and reporting times vary across sites. Network latency and batch processing further disrupt timeliness.

"Bad" Data. Bad data includes but is not limited to erroneous observation data – individual observations that differ from ground truth; "bad" sites – sites that chronically produce erroneous data; and "bad" metadata including incorrect locations and/or incorrect timestamps. Bad data may include items that are not individually considered outliers.

Other Quality Factors. There are many other quality factors including reliability (site, sensor, communication network), timeliness of data, imprecision of data, and imprecision of metadata.

IV. DEFINITIONS AND APPROACH

A. General Definitions

An individual site refers to a fixed-location facility that houses one or multiple sensors that measure conditions. A measurement and associated metadata are referred to as an observation. The set of all sites, represented by S, is the set of sites for which observations are available for a time period and geographic area of interest.

An observation, obs, is represented as a 4-tuple, $obs = \langle s, t, l, v \rangle = \langle obs_s, obs_t, obs_l, obs_v \rangle$ consisting of the site/sensor s, timestamp t, location l (spatial coordinates), and an observed value v. We investigate observations from a single sensor type, so we assume that s identifies both the site and sensor. The set of all observations, represented by O, consists of observations from sites in S over a time-period of interest.

Ground-truth is the exact value of the condition that a given sensor is intended to measure at a given location and time. Ground-truth will rarely be known because of sensor error, estimation error, and high human costs, among other reasons. Human cost is a huge challenge, with agencies struggling to accurately inventory assets and technicians unable to service and maintain all equipment, including situations where they may not even be able to find the equipment.

We wish to evaluate observations to determine if they are erroneous. To do so, we compare observations to estimates of ground-truth. For our purposes, these estimates will be determined via interpolation, which is commonly used in the GIS community, as well as in the weather and road-weather communities.

B. Approach

Identification of Outlyingness and Outliers. We measure outlyingness as the absolute deviation between an observed value and ground truth. Ground truth may not be known, so we estimate outlyingness as the absolute deviation between an observation and modeled ground truth corresponding to the observed value in time and location. Given the degree of outlyingess (exact or estimated), we identify outliers using a threshold. If the degree of outlyingness for an observation meets or exceeds the threshold, then we flag the observation as an outlier. Otherwise, we flag it as an inlier. The degree of outlyingness is more informative than an outlier/inlier label.

Our approach is consistent with general model-based approaches for outlier detection found in Han et al. [7], Tan, Steinbach and Kumar [10] and Aggarwal [11], and follows the general data-mining framework of Train, Test and Evaluate.

C. Interpolation to Model Ground Truth

IDW estimates ground truth as the weighted average of observation values using (geographic) distance from the site for which an observation is to be estimated as the weight, raised to some exponent h. If ground truth is known, a suitable exponent h can be determined to minimize error. Isaaks and Srivastava [41] indicate that if h=0, then the estimate becomes a simple average of all observations, and for large values of h, the estimate tends to the nearest neighboring observation(s). This simple version of IDW does not account for time, so it is assumed that observations fall in temporal proximity.

Least Squares Regression (LSR) estimates observed values using the coordinates of the sites. We only use *x*-*y* coordinates in our experiments for LSR. There could be benefit in using elevation and other variables including time. However, doing so compounds problems related to bad metadata, such as incorrect locations, bad timestamps and inaccurate elevations.

UK estimates observed values using the covariance between sites, the coordinates of the sites, and the observed values. In our experiments, we used a Gaussian covariance function of distance and estimated the related parameters to minimize error relative to ground-truth for our training data using data from the present time window. Refer to Huijbregts and Matheron [42] for further information on UK. We implemented a fitter/solver for the estimation of the covariance function parameters using the Gnu Scientific Library (GSL) non-linear optimization code [43]. Refer to Bohling [44] for additional covariance functions.

These methods can be applied using a restricted radius or a bounding box to alleviate computational challenges and to focus on local trends. Other interpolators could be applied in a similar manner. There are obvious risks in using interpolators. Outliers and erroneous values will have an adverse impact on interpolation, causing poor estimates. Lack of data in proximity to a point to be estimated can also result in a poor estimate. For these reasons, we developed our own robust interpolator in prior work.

D. Our SMART Approach

In prior work, we developed a representative approach for data quality assessment of site-based, spatio-temporal data using what we call Simple Mappings for Approximation and Regression of Time series (SMART) [26-32]. We used the SMART mappings to identify bad (inaccurate) observations and "bad" sites/sensors, so that they can be excluded from display and computation, and to subsequently estimate (interpolate) ground truth.

Site-to-Site Mappings. Let an observation be represented as $obs = \{(t, v): t = time, v = value\}$, pairing the value with the reported time. Let obs_i be the set of observations from site i and obs_j be the set of observations from site j. For a given time radius r we pair the observations from sites i and j as $obs_{pairs_{i,j}} = \{(x, y): (t_1, x) \in obs_i, (t_2, y) \in obs_j, |t_2 - t_1| \leq r\}$. We then define a site-to-site mapping l as a linear function of the x-coordinate (the observed value from site *i*) of the paired observations $obs\_pairs_{i,j}: l_{i,j}(x) = a + bx$. We determine this function to minimize the squared error between the values of the function and the y-coordinates (the observed values from site j) for the paired observations.

We next determine a quadratic estimate q of the squared error of the linear mapping relative to the time offset between the paired observations. We expect an increased squared error for increased time differences. This model estimates the squared error and accounts for time offsets between observations. Our method does not require a complex, data-specific covariance model.

These simple mappings are the core elements of our approach, and we must overcome the potential impact of erroneous data in determining them. LSR suffers from sensitivity to outliers. We use the method from Rousseeuw and Van Driessen to perform Least Trimmed Squares Regression [45]. Least Trimmed Squares determines the least squares fit to a subset of the original data by iteratively removing data furthest from the fit. Before applying least trimmed squares to determine the linear mapping, we select the percentage of data that will be trimmed. We can interpret the trim percentage either as our willingness to accept bad data in our models or our estimate of how much data is bad. We used a trim percentage of 0.1 throughout.

For the quadratic error mappings, we experienced problems with local minima when attempting quadratic least trimmed squares. Instead we group data into intervals, determine the trimmed mean for each group, and then compute the least squares quadratic fit for the (*time difference, trimmed mean*) pairs.

We then check the coefficients and derived measures of the linear and quadratic mappings for outlying values relative to all other mappings. If we find outlying values, we flag the mapping as unusable. For instance, if the axis of symmetry of the quadratic error mapping is an outlier relative to that for another pairing, then there may be a problem with the timestamps of at least one of the two sites.

SMART Interpolator. Our SMART interpolator uses these mappings. Formally: Let S be the set of all sites. Let $s \in S$ be a site for which we are evaluating observations. Let $\langle s_1, \ldots, s_n | s_i \in S, s_i \neq s \rangle$ be the set of sites other than site s. We want to estimate $obs_s(t_s)$, the value of the observation at site s at time t_s using the most recent observations from the other sites relative to time t: (t_i, v_i).

Our SMART interpolator is like IDW, using our quadratic error estimates instead of distance given the time lag between observations and using our SMART linear mappings to yield estimated ground truth producing an estimate. Neither distance nor direction are directly used. The linear mappings and quadratic error estimates account for similarity between sites. No attempt is made to down-weight clustered sites, although there may be benefit in doing so.

We determine the exponent g by minimizing error relative to ground truth, if available, or estimated ground truth. Prior to computing the weighted estimate, we examine the weights and, if necessary, "re-balance" to reduce the potential influence of single sites on the outcome. We found it useful to restrict the maximum relative weight a site can be given to 0.25 to reduce the risk that a bad value from one site will overly influence the resulting average. Rather than take a simple weighted average, we use a trimmed mean to further reduce the influence of outliers.

E. Artificial Data Set

We developed a weather-like phenomenon representing temperature as approximate fractal surfaces produced using the method of Successive Random Addition. For further information on Successive Random Addition, refer to Voss [46], Feder [47], and Barnsley et al. [48]. Fractional Brownian processes were used by Goodchild and Gopal to generate random fields representing mean annual temperature and annual precipitation for the purpose of investigating error in [49]. We used a similar approach to model time series in [50]. A 513x513 approximate fractal surface, $surface(x, y)$, was generated with Hurst Exponent H=0.7 and $\sigma^2 = 1.0$, representing elevation. A 1025x513x513 fractal-like weather pattern, $weather(x, y, t)$, was also generated with Hurst Exponent H=0.7 and $\sigma^2 = 1.0$. The larger x-coordinate allowed us to simulate motion/flow. We generated one surface and eight weather patterns, allowing us to train on one weather pattern and test on those remaining.

We generated time series of "ground truth" data by combining the surface data with the weather data, a periodic effect and a north-south effect to simulate a weather-like phenomenon like the diurnal effect and general north-south variation in the Northern Hemisphere respectively. We added the weather data as is, with varying offsets in the x-coordinate to represent a west to east flow in the weather pattern. The surface value is subtracted so that low points are "warmer" than high points. The periodic effect represents warming during the day and cooling at night. The north-south effect yields warmer points to the south and cooler points to the "north". Our approach yields a time series of length n=513 for each (x, y) on the 513x513 surface.

We selected 250 "sites" using random uniform x-y (spatial) coordinates. For each site we assigned a reporting pattern with a random frequency and offset. We added errors to the observations from 25 sites via: random noise added to ground truth (NOISE), rounding of ground truth (ROUNDING), replacement of ground truth with a constant value (CONSTANT), replacement with random bad values with varying probabilities (RANDOMBAD), or negation of ground truth. The remaining 225 sites were left error-free.

V. EVALUATION

We evaluated the performance of the various interpolators including our SMART Method in-depth, in terms of computation and ability to identify bad data. We compared our SMART method, IDW, LSR, UK and OK. We measured performance and scalability using run-time in milliseconds. We measured accuracy using mean-squared-error (MSE) between estimated and known ground-truth. We compared means using t-tests when multiple runs were available. We used Area Under the ROC Curve (AUROC) analysis to evaluate accuracy of outlier classification given varying "threshold" values for outlier/inlier determination.

We analyzed our artificial data set, MADIS air temperature for Northern California from December 2015, MADIS air temperature for Montana from January 2017, and Average Daily USGS Streamflow for Montana from 2015, 2016, 2017.

A. Evaluation Using our Artificial Data Set

We performed an in-depth comparison of the various algorithms using our artificial data set. We enhanced the standard algorithms by randomly choosing neighboring sites using set inclusion percentages (0.1, 0.2, 0.3, …, 0.9, 1.0). For instance, a 0.9 inclusion percentage corresponds to selecting neighboring sites individually with 0.9 inclusion / 0.1 exclusion probability. We varied the radius (50, 75, 100, …, 175, 200) over which sites were included relative to the location of the site whose observation we were testing. We repeated this procedure 10 times for each parameter combination (inclusion percent and radius) and used the median of the resulting estimates as the estimate for that parameter combination. By randomly holding out sites, bad data will be held out in some of the resulting combinations. By taking the median of the results, we eliminate the extreme estimates, particularly those impacted by bad data, and ideally determine a robust estimate.

We ran the methods in aggregate over the eight time periods spanning 512 time units. For each time period, there were 37,293 observations total from the 250 sites. We iterated through the observations in order by time and estimated ground truth for each observation as if computing in real time as the observations become known. Only observations that occurred at the same time as or prior to each observation were used for prediction, simulating real-time operation of the system. We averaged the MSE and run time for each configuration (inclusion radius and inclusion percent).

We compared the results of the various runs of the methods. The run time for the SMART method was 6336.6 ms, and the MSE was 0.1026. The SMART method was comparable in run time to IDW, but the accuracy achieved was far better than for any of the other methods.

We measured the ability of each method to distinguish increasing percentages of the bad data from good data using an AUROC analysis. True outliers were defined as data that differs from ground-truth – i.e., data that was modified to be erroneous. Predicted outliers were data that differed from estimated ground truth by a given threshold. We varied thresholds for outlier/inlier cutoffs and compared results with the actual labels identifying whether the data was truly an outlier or inlier. The AUROC (area under the ROC curve) values are shown in Table I. The AUROC values show better discriminative power for the SMART method versus the other methods. No method will be perfect in identifying all errors. Some errors are small and impossible to distinguish from interpolation error. Known ground truth and known error from ground truth yields perfect labels.

TABLE I. AUROC VALUES FOR ARTIFICIAL DATASET

Method	SMART	UK	LSR	IDW
AUROC	0.827	0.740	0.739	0.708

Our SMART method's computation time is comparable to IDW and is far better than LSR and UK, but we still should account for the preprocessing computation time required for determining the linear mappings and quadratic error functions. The overall amount of preprocessing time required to determine the linear mappings and quadratic error functions was comparable to run time required for UK. This was encouraging. Generation of the mappings will be done as an offline, batch process, so the observed time required is still within reason to help facilitate the faster and more accurate, online process. Additional benefits, such as identification of bad sites and bad metadata, come from these mappings, further justifying the effort required. Optimization can reduce the overall time needed to compute the mappings. The benefits and potential to improve the run time outweigh the amount of required preprocessing time.

B. December 2015 MADIS California Data

We analyzed Northern California December 2015 ambient air temperature data from the MADIS Mesonet subset. We used a bounding box defined by $38.5° \leq latitude \leq 42.5°$ and $-124.5° \leq longitude \leq -119.5°$, yielding 888 sites. We excluded observations that failed the MADIS Level 1 Quality Control Check. This range check restricts observations in degrees Fahrenheit to the interval [-60°F,130°F]. Many values failing this check fall far outside the range and can have a dramatic impact on the interpolation methods. Our SMART method performs very well in the presence of extreme bad data, and it would have easily out-performed the other methods in the presence of the range-check failed data.

There were over 2 million observations. MADIS flagged 73.5% of these observations as "verified" / V, slightly less than 4% as "questioned" / Q, and 22.5% as "screened" / S, indicating that it had passed the MADIS Level 1 and Level 2 quality checks, but that the Level 3 quality checks had not been applied.

Training. Verified (V) observations from the first week in December 2015 were used to train all methods, including our SMART method. In the absence of range-failed data, the "enhanced" (iterated subset) versions of the other algorithms showed little improvement in accuracy while consuming excessive computation time, particularly "enhanced" UK. In some cases, it would have taken days to compute results. Because of this, we used the methods directly, without enhancement. We also tested OK (refer to Bailey and Gatrell [51] for further information). Since we do not know "ground truth" for this data, the verified data is the closest to ground truth. We trained all methods on this data to MSE of predicted versus actual. We used a 50-mile inclusion radius due to the density of sites to avoid excessive computation time for the kriging approaches.

The SMART mapping coefficients and derived values were examined for outliers, and ranges were determined for valid mappings. If any coefficient or derived value for a given SMART mapping fell outside these ranges, then the SMART mapping was considered bad, and that mapping was not used for predictions.

Our SMART method produced significantly better results than all other methods for the training data in terms of estimation of ground truth measured by MSE, as shown in Table II. A paired, one-sided t-test was used for significance testing using paired squared errors from predicted values. Only the verified (V) data was used in this comparison since it best approximates ground truth. The SMART method was compared pairwise with the other methods and results were aggregated over instances where both methods produced predictions.

TABLE II. MSE FOR MADIS CALIFORNIA TRAINING DATA

Method	**MSE**	**Method**	**MSE**
SMART	2.8322	IDW	7.6212
SMART	2.8322	LSR	17.1446
SMART	2.8046	OK	18.4989
SMART	2.8046	UK	16.5289

Testing. Testing was conducted using all data from the entire month of December 2015, minus the range-check-failed data. We computed the MSE for the verified (V) data since it best represents ground truth, but all observations were used in making estimates. The testing results indicate the robustness of methods in the presence of bad data. In comparisons across all other methods, the SMART method significantly out-performed all other methods in terms of MSE, as shown in Table III.

We conducted an AUROC analysis to compare classification ability of the methods based on the MADIS quality control flags. We considered the following flags from MADIS to be good/inlier data: V/verified, S/screened,

good. The Q/questioned, was treated as bad/outlier data. Recall that we excluded the observations having a QC flag of X, those that failed the range test, from our evaluation. Even if we accept the MADIS quality control flags as being correct, and we do not, this approach is problematic. The MADIS QC flag S corresponds to data for which not all the QC checks have been run. While this data had not failed any quality control checks that have been applied, it possibly would have failed the higher-level checks.

In terms of AUROC, IDW, LSR and SMART were comparable, with IDW finishing slightly ahead, as shown in Table IV. While these AUROC values seem reasonable, they are affected by incorrect outlier/inlier labels, and our SMART method suffers the greatest impact because the distance-based methods approximate the MADIS Level 3 quality control check. OK and UK fall short because they fail to make predictions for many observations.

TABLE III. MSE FOR MADIS CALIFORNIA TESTING DATA

Method	MSE	Method	MSE
SMART	4.4611	IDW	9.1306
SMART	4.4611	LSR	16.5223
SMART	4.3360	OK	16.0868
SMART	4.3360	UK	14.2086

TABLE IV. AUROC FOR MADIS CALIFORNIA TESTING DATA

	AUROC
IDW	0.7906
LSR	0.7578
SMART	0.7317
OK	0.6458
UK	0.6062

C. December 2017 MADIS Montana Data

We investigated ambient air temperature for Western Montana / Northern Idaho from the MADIS Mesonet and the MADIS HFMetar subset in January 2017. We added the HFMetar data set to account for aviation AWOS/ASOS sites that had previously been included in the Mesonet data set. We used a bounding box defined by $44° \leq latitude \leq 49°$ and $-116° \leq longitude \leq -110°$, resulting in observations from 497 sites. This bounding box is comparable in size to the one used for Northern California, although the density of sites is less. We excluded observations that failed the MADIS Level 1 Quality Control Check.

All total there were over 1 million observations. MADIS flagged 71.2% of these observations as "verified" / V; 10.3% of as "screened" / S, indicating that they had passed the MADIS Level 1 and Level 2 quality checks, but that the Level 3 quality checks had not been applied; and a relatively large 18.5% of the data as "questioned" / Q. This is over four times the percentage of questioned data as there was for the California data set.

Training. Verified (V) observations from the first week in January 2017 were used to train all methods, including our SMART method. We used a 100-mile inclusion radius due to a low density of the Montana/Idaho sites. The SMART mapping coefficients and derived values were examined for outliers, and bad mappings were identified as any mapping associated with such values. The quality of the mappings as measured by MSE was noticeably less than that for the Northern California data set. We found problems with many of the timestamps in this data set. Recognizing that much of the Idaho data comes from the Pacific Time Zone while the Montana data comes from the Mountain Time Zone, there appeared to be many sites for which the conversion to UTC time was not consistent. The Northern California data all falls within Pacific Time, and we did not see this problem in that data set. In terms of MSE, the SMART method produced significantly better results than each of the other methods for the training data, as shown in Table V.

TABLE V. MSE FOR MADIS MONTANA TRAINING DATA

Method	MSE	Method	MSE
SMART	8.1513	IDW	16.7217
SMART	8.1513	LSR	29.8039
SMART	10.6726	OK	47.0028
SMART	10.6726	UK	33.9863

Testing. Testing was conducted using data from the remainder of January 2017. All data was used for this test except for the observations that failed the MADIS Level 1 range test. The SMART method significantly outperformed all other methods in terms of MSE, as shown in Table VI.

TABLE VI. MSE FOR MADIS MONTANA TESTING DATA

Method	MSE	Method	MSE
SMART	21.4714	IDW	38.3208
SMART	21.4714	LSR	38.5063
SMART	23.1496	OK	50.5078
SMART	23.1496	UK	36.6762

We conducted an AUROC analysis to test classification ability based on the MADIS quality control flags in the same way as described for the Northern California data set in the previous section. As noted in that section, many of the MADIS QC flags are incorrect. In terms of Area Under the ROC curve, LSR, IDW and SMART were comparable, with LSR finishing ahead, as shown in Table VII. These AUROC values are less than those for the Northern California data set at least in part because all methods adversely affected by incorrect outlier/inlier labels.

TABLE VII. AUROC FOR MADIS MONTANA TESTING DATA

Method	LSR	IDW	SMART	OK	UK
AUROC	0.6900	0.6697	0.6393	0.5432	0.5476

This data set includes a large percentage of observations (18.5%) that are flagged as "questionable" by MADIS. These were considered "bad" / outliers for the purposes of our analysis. It also includes a large percentage (10.3%) that are flagged as "screened" by MADIS, indicating that not all QC checks have been conducted. These are considered "good" / inliers for our analysis.

There were many observations flagged as "questionable" / outliers in the HFMetar subset that should have been

flagged as "good" / inliers. This data alone accounts for most of the questionable data in the data set. Aviation weather sites are well-maintained and regularly calibrated, so it is hard to believe that these sites would produce data that is entirely bad. We checked this data against predicted values, as well as neighboring sites, and it was very close, so it is unclear why the data was labeled as questionable.

Numerous sites were flagged by our SMART method as "bad" and all observations from those sites were labeled as bad. MADIS flagged some observations from these sites as good when they were close to predicted values. In some cases, this may have been reasonable, but in others it was a random occurrence. There were some sites that produced bad data for the training period but then produced good data for at least a portion of the test period. One could argue that for such sites all associated observations should be questioned. If a site was identified as bad by the SMART method, then the V and S observations would adversely impact the SMART method in the AUROC analysis. The chance situations in which the other methods came close to the "good" values and far from the "bad" values improved their performance.

D. December 2015-2017 USGS Streamflow Data

Mean daily streamflow (ft^3/sec) was downloaded for all sites in Montana from the USGS [52] for every day from January 1st, 2015 through April 24th, 2017. There were 145 sites having data than spanned this period, and these sites were analyzed. This data set is far different from the air temperature data used for prior analysis. Since daily averages were used, there is no visible diurnal effect. There is a seasonal effect which varies with elevation and location relative to watersheds. Due to the dramatic fluctuations that occur in this data during times of peak runoff, the base-10 logarithm of the data was used for analysis.

This data set includes quality flags. Daily values are flagged as "A", approved for publication, and "P", provisional and subject to revision. Values may further be flagged as "e" for estimated. Values transition from provisional to approved after more extensive testing is conducted, so provisional values aren't necessarily bad. These flags were of limited use to us and we did not use them for analysis. We treated the data as being all good and subsequently introduced errors into some of the observations, making them known bad. There were 122,380 total observations.

Training. All data from 2015 was used to train all methods, including our SMART method. We assume this data, which was mostly "approved", to be ground truth. We trained over this data to minimize MSE of predicted versus actual. We used a 200-mile inclusion radius. The SMART mapping coefficients and derived values were examined for outliers. If any coefficient or derived value for a given SMART mapping was an outlier, then the SMART mapping was considered bad, and it wasn't used for predictions. In terms of MSE, the SMART method produced significantly better results than the other methods for the training data, as shown in Table VIII.

Testing. Testing was conducted using the 2016-2017 data. The SMART method significantly outperformed all other methods in terms of MSE, as shown in Table IX.

TABLE VIII. MSE FOR USGS TRAINING DATA

Method	MSE	Method	MSE
SMART	0.0174	IDW	0.8751
SMART	0.0174	LSR	0.9611
SMART	0.0174	OK	0.9431
SMART	0.0174	UK	0.9617

TABLE IX. MSE FOR USGS TESTING DATA (NO ERRORS)

Method	MSE	Method	MSE
SMART	0.0429	IDW	0.9031
SMART	0.0429	LSR	0.9869
SMART	0.0429	OK	0.9755
SMART	0.0429	UK	0.9874

Testing was then conducted using the 2016-2017 data, with errors introduced into 10% of the observations. A random normal value with mean zero and standard deviation one was added to each of the observations in the 10% group. The MSE was computed relative to the known, original observations which represent ground truth, and all observations (including bad observations) were used in making estimates. The testing results help to indicate the robustness of methods in the presence of bad data. The SMART method significantly outperformed all other methods in terms of MSE, as shown in Table X.

TABLE X. MSE FOR USGS TESTING DATA (WITH ERRORS)

Method	MSE	Method	MSE
SMART	0.0453	IDW	0.9103
SMART	0.0453	LSR	0.9907
SMART	0.0453	OK	0.9776
SMART	0.0453	UK	0.9914

We conducted an AUROC analysis to test the methods on classification ability based on whether observations had been altered to be erroneous by our process of randomly selecting 10% of the observations and adding a normal random variable with mean 0 and standard deviation 1 to those observations. The altered observations were labeled "bad"/outlier and the unaltered observations were labeled as "good"/inlier. Our SMART method performed far better than all the other methods, achieving an AUROC value of 0.8722, as shown in Table XI. The other methods had values between 0.6 and 0.63.

TABLE XI. AUROC VALUES FOR USGS TESTING DATA

Method	SMART	IDW	OK	UK	LSR
AUROC	0.8722	0.6241	0.6136	0.6046	0.6031

E. Evaluation Summary

For all four data sets and for every training and testing instance compared, our SMART method performed significantly better in terms of accuracy (MSE) than all other methods. Its computational performance was competitive

even though no effort was made to optimize it. For the two MADIS data sets, its performance for AUROC analysis of classification and discrimination capability showed it to be competitive with the best of the other methods. This comparison and evaluation made use of MADIS data quality labels for which we have found numerous problems. As such, all methods underperformed, and the SMART method was penalized most by mislabeling. For the other two data sets (artificial and USGS) in which ground truth is known or assumed and errors were introduced relative to ground truth, the SMART method outperformed the other methods by a wide margin. This further supports our assertions regarding the impact of bad labels on the MADIS data, and the need for better methods and benchmark data sets for data quality assessment.

OK and UK both failed to produce estimates for many observations, likely due to singular matrices. They were not competitive in terms of run time and their accuracy was no better than the other methods. UK and LSR are prone to occasional very large errors if the predicted surface slopes in an extreme manner.

Our SMART method identifies "bad sites" that chronically produce bad data and does not use data from these sites in estimating ground truth for other sites. Similarly, data from these "bad sites" is labeled as all bad. The SMART method falls short in cases where a site exhibits chronic behavior during training but recovers to produce good data during a testing period.

The USGS streamflow data exhibits correlation between sites, but the correlation corresponds to sites close to each other and in the same river/stream. Correlation will not necessarily be high for sites that are close but in different rivers. For rivers that have dams and other features that may influence streamflow in unusual ways, sensors will be correlated on each side of such features, but not as much on opposite sites, and certainly not as much with sites on rivers that do not have similar features.

The SMART method identifies like sites, yielding better correlations. IDW and LSR will not perform well in this circumstance. And, the kriging methods will not perform well either if a stationary, isotropic covariance function is used. Such an assumption is typical, and we used this assumption in determining the covariance matrices for the kriging tests.

VI. Analysis of Mislocated Sites

Bad metadata presents significant challenges regarding data quality assessment. As discussed earlier in this paper, many if not most techniques for spatial-temporal quality assessment depend on distance and time directly and/or related assumptions. If timestamps on some data are incorrect, then that data will be compared against data from other time periods, times in which conditions may be dramatically different. If location metadata is incorrect, specifically if a site is mis-located, then it may be compared against sites that are far away, and quality assessment will suffer. Conditions may vary dramatically by location, even if timestamps are correct. This causes a chicken-egg problem: to assess the quality of data, we need quality data for which the quality has been determined via quality assessment. How do we find and filter the bad data such as that with bad location metadata amid other data quality issues? In this section, we present a new analysis of this problem.

A. Gibson near Castella

For several years, we were aware of a mis-located site in the MADIS feed that was displayed in our WeatherShare system. WeatherShare users identified this site as mis-located in correspondence with us. Site GISC1 (Gibson near Castella) was mis-located at latitude 38.56556° N, longitude -121.485° W in downtown Sacramento prior to a correction that was made sometime in 2016. Note that MADIS obtained data for GISC1 from the National Weather Service's Hydrometeorological Automated Data System (HADS) system. Subsequently the location of GISC1 was corrected in the MADIS feed to latitude 41.022° N, longitude -122.399° W, 175 miles to the north near the Caltrans Gibson Maintenance yard and near the town of Castella. This relocation makes sense given the name of the site and the error reports from our WeatherShare users. There was no apparent indication why/how this site was mis-located, and we are unsure of how it was relocated. It may have been initially assigned the coordinates of another site.

We didn't have a mechanism for dealing with issues like this. We could have taken our users' word and manually relocated the site within our system by changing the latitude and longitude to what they reported. But, how could we confirm they were correct? And, if we manually changed the location, then we would also need a mechanism to detect if the location was subsequently modified in the feed, perhaps giving a better indication of the true location. We could contact the provider and ask them to correct the situation, but they too may not know the true location of the site. Even in instances where errors were known, it has taken providers years to address data quality issues we have informed them of. Instead, we chose to suppress the display of this site, and implemented a mechanism allowing us to manually select and suppress the display of any site. Of course, this out of sight (no pun intended) out of mind approach was not perfect either because again, the location of the site might be corrected in the feed at some point. Subsequently, we made a choice to again display all data with the caveat that users would have to decide for themselves what data was good and what was bad. That approach is less than ideal and could give the perception of poor quality of a system overall.

At the same time, we wondered if the mis-location of GISC1 was isolated or if there were more mis-located sites. Given our experience with other data quality issues in this data set, we suspected the latter. But the reality was that we had no way of knowing for sure in the absence of working directly with owners and operators of the equipment in the field and/or conducting site visits. That was well outside the scope of our work and would otherwise have been cost-prohibitive and infeasible. A more practical question was whether we could find such errors automatically. We have

spent significant time since then investigating this challenging and interesting problem.

In this section, we present a new analysis of MADIS ambient air temperature data from sites located in Northern California between 2014 and 2019. We look in retrospect at the 2014 data to see what more we could have done at that time to identify such problems and to assess the state of the problem at that time. We look at temperatures from the month of January, as considerable variation in temperature occurs in that month in conjunction with the bad weather season in Northern California. We selected sites that reported at least once within 90% of the 15-minute intervals during the given month. Overall there were 575 sites that met these criteria for 2014. We used all data regardless of MADIS quality control assessment.

B. Relocated Sites in the MADIS Dataset

One possible indication of erroneous location metadata is the subsequent revision of site locations in the feed. We investigated the 2014 data versus the subsequent 2015 through 2019 data for each site and identified all changes in location. We found that 5 sites were subsequently relocated by more than 20 miles from the locations provided in the 2014 data. GISC1 was relocated furthest at 176.6 miles, matching our earlier observations. Another site, KLHM, the Lincoln Regional Airport, was relocated 37.6 miles from its 2014 location. The three other sites were unfamiliar to us and trying to confirm their true locations could be challenging. See Table X.

TABLE X. 2014 SITES RELOCATED 20 MILES OR GREATER

Site	Year of Change	Distance in miles from Original Location
GISC1	2017	176.6
TS389	2018	60.0
KLHM	2018	37.6
TT109	2016	33.2
SNWC1	2016	27.7

Overall, 109 of the 575 sites were relocated at least once between 2015 and 2019 relative to their 2014 locations. While this may seem like a large proportion, and it is, most of the changes were relatively small. 81 sites were relocated less than a mile from their 2014 locations. Perhaps greater precision was used in specifying their locations: for instance, using two or more digits beyond the decimal for specification of latitude and longitude versus one digit. That could account for changes in location of several miles, and there were 93 sites overall that were relocated 5 miles or less from their 2014 locations. That leaves only 10 sites relocated by between 5 and 20 miles, plus the five sites shown in Table X that were relocated by 20 miles or greater. See Table XI. As such, and if the relocations are correct, we could say that 15 of the 575 sites were mis-located in the 2014 data feed, but we truly do not know. In fact, we found some sites that were relocated multiple times including one site, TR180, that was relocated in 2018 to a point 24.5 miles from its 2014 location and then relocated again in 2019 to a point 0.015 miles from its 2014 location. It was moved some distance away and then move back to almost the same original location. Which location, if any of the three, was right?

TABLE XI. 2014 SITE RELOCATIONS COUNTS BY DISTANCE

Range	Count
0 to 1	81
1 to 2	6
2 to 5	7
5 to 10	2
10 to 20	8
20 to 50	3
50 to 100	1
100+	1
TOTAL	**109**

C. Identifying Mis-located Sites with our SMART Method

As discussed earlier, our SMART mappings can provide a robust, consistent measure of dissimilarity between sites in the presence of bad data. We did not filter the original data other than by time and location. We did not use the MADIS quality control flags to filter at all. Bad data is certainly included. Because of this, measures of correlation or covariance would be adversely affected by bad data. We have found that the mean-squared-error (MSE) of the SMART mappings provides a robust alternative for measuring dissimilarity. We also expect to find, in general, that near sites are more closely related than far sites (Tobler's First Law of Geography). This relationship can be exploited to identify mis-located sites – at least those that are severely mis-located.

Before proceeding, we needed to be cognizant of the challenges presented earlier in this paper, as they would certainly have an impact on the results. In order to overcome some of these challenges, particularly bad data and challenges with non-uniform time reporting, we chose somewhat loose parameters for our SMART mappings: a time radius of 90 minutes for pairing observations, a 10% cutoff for trimmed least squares regression, and a 10% trim percentage for computing trimmed means. While these choices help to overcome the stated challenges, they may also blur the relationships between sites, which can cause challenges in the presence of non-isotropic covariance. In future work we plan to investigate parameter selection further.

Now we look at the relationship between distance and MSE of SMART mapping from other sites to GISC1 using the 2014 data, including the incorrect location of GISC1 in the 2014 data. Again, we expect near sites to be more closely related (low MSE) than far sites. But we find instead that sites falling over 150 miles away have the lowest MSE for the SMART mappings. See Figure 1.

Next, we look at the same plot for the 2014 GISC1 data, but with the location corrected to the 2017 location. See Figure 2. In this figure, sites nearest the corrected location have the lowest MSE values. And, there appears to be an apparent, positive trend in which MSE increases by distance. But there is also a lot of variation. This variation is

most likely attributable to non-isotropic covariance and bad data. Tightening the parameters used for the SMART mappings might help to reduce this variation, and we intend to investigate this in future research. Regardless, the plot does appear to confirm that GISC1 was relocated to the corrected location or at least near the correct location.

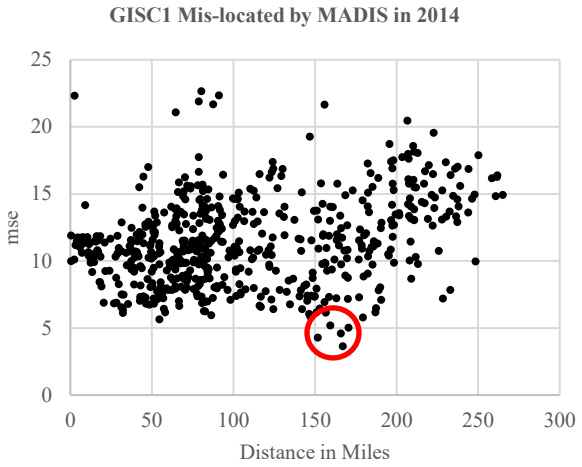


Figure 1: GISC1 (Incorrect Location) Distance versus MSE of SMART Mapping by Site

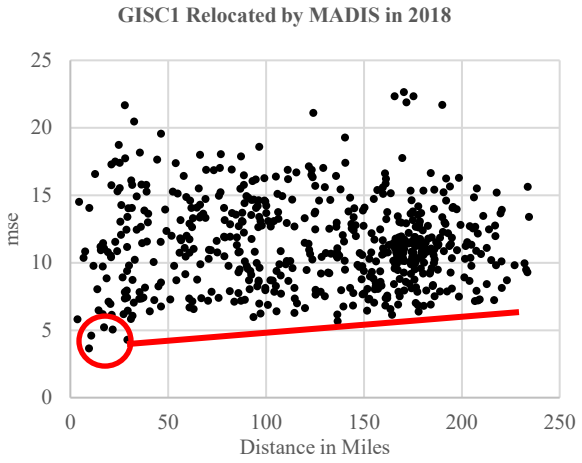


Figure 2: GISC1 (Corrected Location) Distance versus MSE of SMART Mapping by Site

Next, we look at site KLHM, the Lincoln Regional Airport. Recall that the Lincoln Regional Airport site was relocated subsequent to 2014 by 37.6 miles in the MADIS feed. In plots of distance versus MSE, we checked to see if the same relationships hold. See Figure 3 and Figure 4. Again, we see that the sites having the lowest MSE for the SMART mappings fall approximately the same distance from the mis-location as the subsequent re-location. And, we see a more prominent positive trend when the site is relocated. In this case, it is easier to confirm the correct location since the location of the Lincoln Regional Airport is known. Still, an airport occupies a lot of space, and the exact location of the weather sensor at the airport is not known to us with certainty.

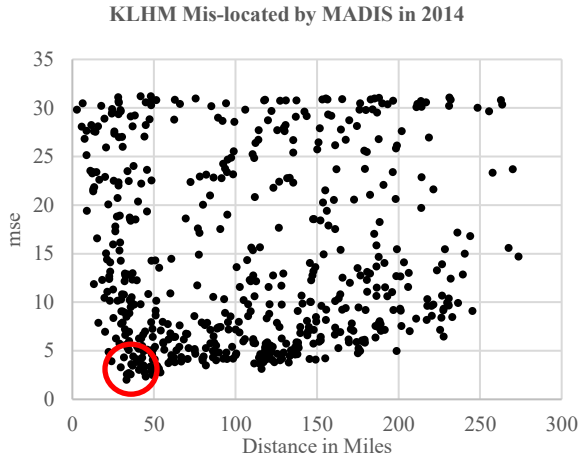


Figure 3: KLHM (Incorrect Location) Distance versus MSE of SMART Mapping by Site

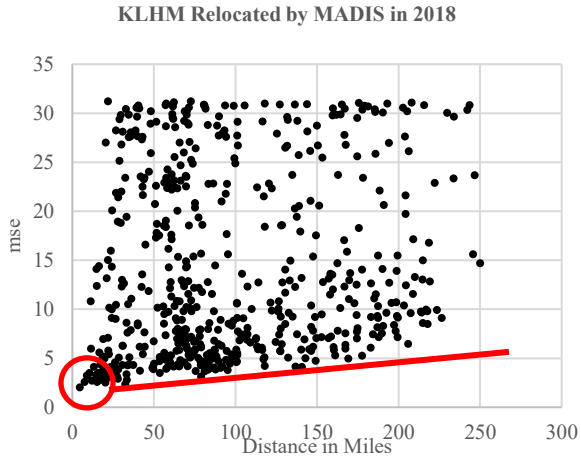


Figure 4: KLHM (Corrected Location) Distance versus MSE of SMART Mapping by Site

Before proceeding to develop a more formal method to identify situations where sites may be mis-located like GISC1 and KLHM in the 2014 MADIS data, we will artificially mis-locate several sites for which we know the correct locations, and we will see if these relationships hold. For this we will use the weather station at the Redding Airport, KRDD, and the weather station at Sacramento International Airport, KSMF. Both sites do appear to be correctly located in the 2014 MADIS data with the caveat that their precise location at each airport is unknown to us and the specification of location may lack some precision.

For the Redding Airport, KRDD, we artificially relocate the site 130 miles to the south. We see the same signature patterns in the corresponding plots. See Figure 5 and Figure

6. When correctly located, sites having the lowest MSE for the SMART mappings fall closest to the site and there is a general upward trend in the data. When the site is incorrectly located, sites have the lowest MSE for the SMART mappings fall at a distance corresponding to the distance between the correct and incorrect locations.

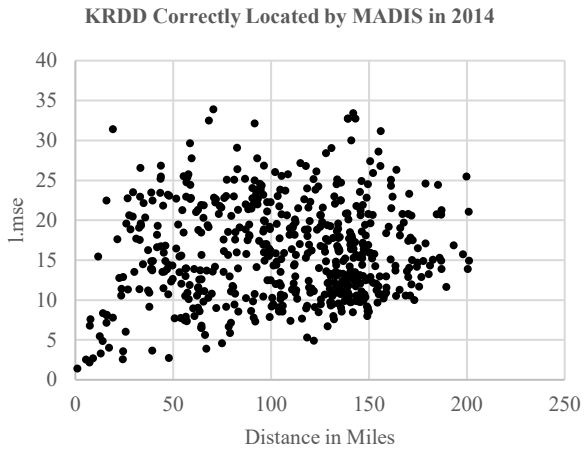


Figure 5: KRDD (Correct Location) Distance versus MSE of SMART Mapping by Site

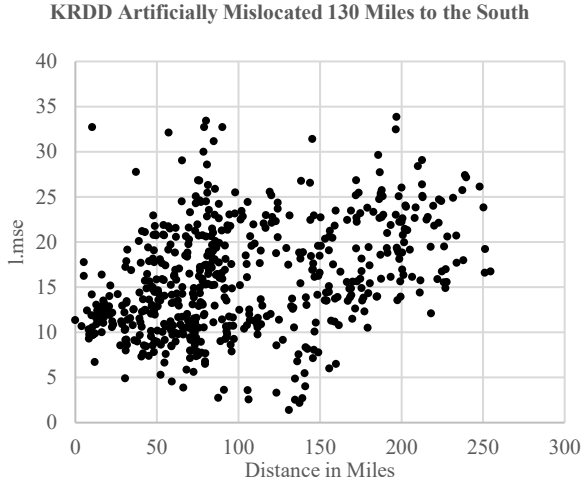


Figure 6: KRDD (Incorrect Location 130 miles south) Distance versus MSE of SMART Mapping by Site

For the Sacramento International Airport, KSMF, we artificially relocate the site 175 miles to the north. (This is similar but in the opposite direction of the mis-location of GISC1.) Again we see the signature patterns in Figure 7 and Figure 8.

Site TS389 was relocated in the 2018 MADIS feed 60 miles from its 2014 location. Given the evidence we presented above, it should be apparent from similar plots if this relocation was correct. As we just saw with the artificial relocations of KRDD and KSMF, we can also spot situations in which a site was incorrectly re-located. This may be the case with site TS389. See Figure 9 and Figure 10. It appears that the 2014 location was correct. And, it appears that the 2018 relocation of the site was incorrect. Realize though that these plots show relationships for 2014 data. It could be the case that this site truly was moved in 2018 or that the old site ceased operation and a new site was given the same name. This may seem strange, but given vague naming on some of these sites, it is possible. While not the case here, there is certainly the potential for "mobile" site data to be incorporated into the feed, in which a portable sensor suite is moved from location to location as needed. Further challenges would occur in assessing constantly moving sites such as vehicles equipped with weather sensors. Such data is being collected by department of transportations and others, and that data is being incorporated into data sets such as MADIS.

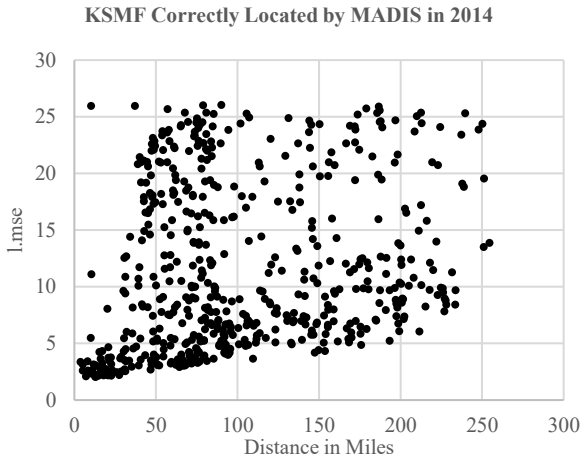


Figure 7: KSMF (Correct Location) Distance versus MSE of SMART Mapping by Site

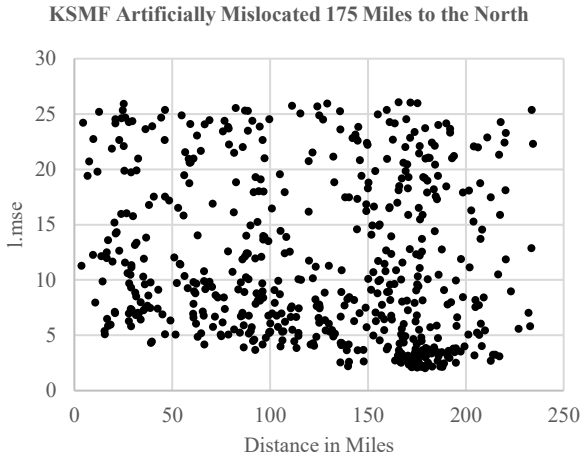


Figure 8: KRDD (Incorrect Location 175 miles north) Distance versus MSE of SMART Mapping by Site

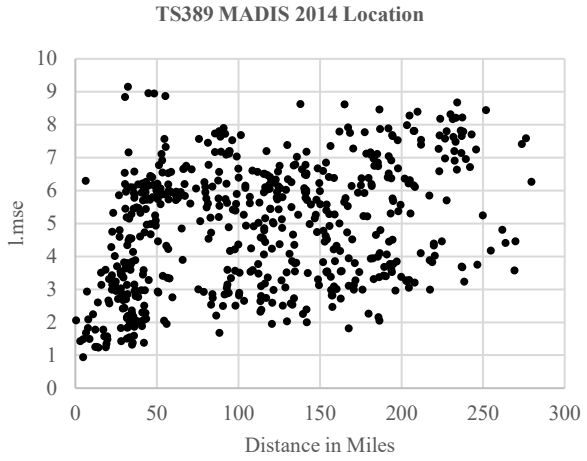


Figure 9: TS389 (Likely Correct Location) Distance versus MSE of SMART Mapping by Site

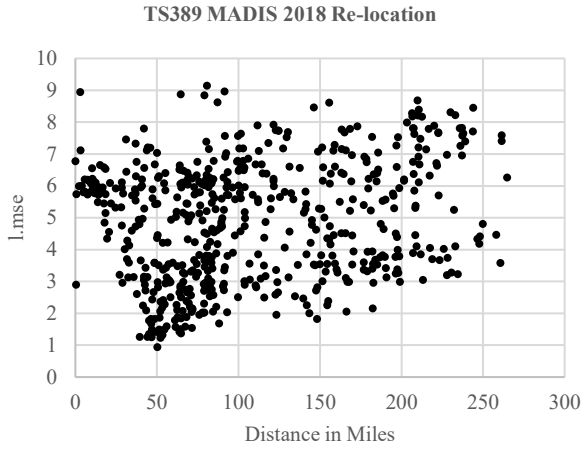


Figure 10: TS389 (Likely Incorrect 2018 Location for 2014) Distance versus MSE of SMART Mapping by Site

Site TT109 was relocated by MADIS in 2016 to a point 33.2 miles from its 2014 location. The plots of the 2014 data relative to these locations do appear to show that the relocation was correct. We omit the plots here for the sake of brevity. However, the plots for SNWC1, which was relocated by MADIS in 2016 to a point 27.7 miles from its 2014 location, are inconclusive. It may be the case that this site has other issues, including chronically bad data, that make it difficult to compare against neighbors.

While these signature patterns seem apparent visually in the plots in most cases, they can be challenging to identify in an automated fashion because of error and variation. We tried a few approaches including using the slope of the best-fit regression line, Spearman's Correlation Coefficient for the ranks of distances versus MSE and found that neither provided a reliable mechanism for identifying situations like those shown above. Instead, we developed a nearest neighbor approach that shows promise.

Let D_k be the set of nearest k neighbors to the site in question by distance. Let L_k be the set of nearest k neighbors by MSE from SMART mappings to the site in question. Then let $J_k = |D_k \cap L_k|/|D_k \cup L_k|$. This is the Jaccard Index, which measures the amount of overlap in the two sets. In our situation, it measures the amount of overlap between the nearest sites in terms of distance and the nearest sites in terms of MSE for SMART mappings. Intuitively, sites that are correctly located should have a high Jaccard Index and sites that are incorrectly located should have a low Jaccard Index. But the selection of *k*, the number of neighbors, could be tricky in the presence of errors and variation. For this reason, we compute $M_L = \max(J_k)$ for $k \in \{5, 10, 15, 20, 25, 30\}$, taking the maximum value as our measure of overlap. The rationale for doing this is that variation and errors in data, including other mis-located sites, may result in low, inaccurate values for small numbers of neighbors. High values of k will be influenced by variation in the data and eventually by the inclusion of most of the sites. Taking the max of the Jaccard index values for the given values of k helps to mitigate these issues.

For the mis-located 2014 location of GISC1, we get M_L = max(0.0, 0.0, 0.0, 0.0, 0.0, 0.0) = 0.0. There is no overlap between the sets of neighbors. For the relocated location of GISC1 we get M_L = max(0.11, 0.18, 0.15, 0.14, 0.14, 0.15) = 0.18. While this value is not close to the maximum possible value for the Jaccard Index of 1, it is an improvement over no overlap. The low value might be explained by significant variation in proximity to the GISC1 site as well as other erroneous data that is nearby. GISC1 sits in the mountainous area north of Redding along the Sacramento River. There is a lot of variation in terrain and variation in weather in that area, particularly during the bad weather season.

For the mis-located 2014 location of KLHM, we also get M_L = max(0.0, 0.0, 0.0, 0.0, 0.0, 0.0) = 0.0, indicating no overlap. For the relocated location of KLHM we get M_L = (0.25, 0.18, 0.2, 0.29, 0.32, 0.33) = 0.33. Relocating the site shows improvement from zero overlap and helps to confirm that the site was correctly relocated.

Given that we know that the GISC1 and KLHM sites were mis-located and we know that they were re-located at, or at least close to their correct locations, we can make such comparisons and see if there is improvement. But when we don't know the correct location for a site, we are left with just the original M_L value. If that value is low or especially if it is zero, then we might suspect that a site is mis-located. But there could be other problems including that a site is producing chronically bad data and there is no relationship between it and any other sites. (We address that problem in separate work.) To address this, we can gather further evidence if we can find prospective locations for which the M_L value is greater or even optimal relative to a set of candidate locations. We can do this using a grid search to identify candidate locations for sites suspected to be mis-located. This leads to the following general logic for identifying sites that we believe are mis-located:

IF the M_L value for the original location is *low* AND the M_L value for the optimal location is *high* AND the optimal location is *far* from the original location, THEN the site may be mis-located, and it should be investigated further.

We realize that this logic is vague, and for the purposes of this paper we will leave it vague. We identified such relationships by manual inspection of the ML values and distances to optimal locations. This process could certainly be automated, but we save doing so for future work.

Now we turn our attention to identifying sites that we suspect are mis-located but that MADIS did not subsequently re-locate. We use the logic above to identify these. We find sites for which there is little or no overlap between nearest sites in terms of distance and MSE of SMART mappings, we identify optimal locations for re-locating those sites, and if the distance between the locations is large, we inspect them further.

For site BUPC1 and its location specified in the MADIS feed, we get M_L = max(0.0, 0.0, 0.0, 0.0, 0.0, 0.02) = 0.02, a small value. The optimal candidate re-location point yields a value of M_L = max(0.11, 0.25, 0.43, 0.38, 0.28, 0.28) = 0.43, a relatively high value. This is a dramatic improvement and deserves further investigation, so we created plots for the original location and the possible (optimal) re-location point. See Figure 11 and Figure 12. These plots do appear to show that site BUPC1 is mis-located relative to the 2014 MADIS data.

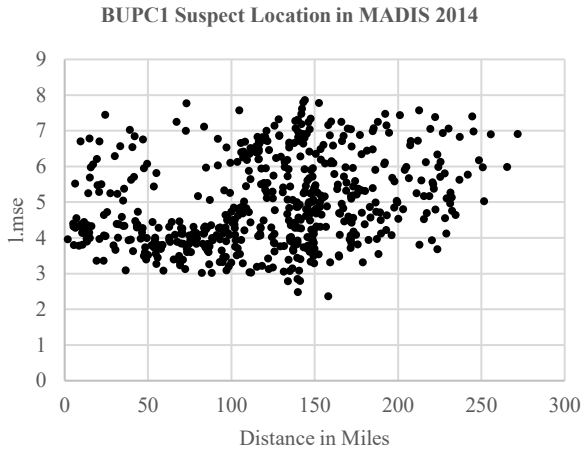


Figure 11: BUPC1 (Likely Incorrect Location) Distance versus MSE of SMART Mapping by Site

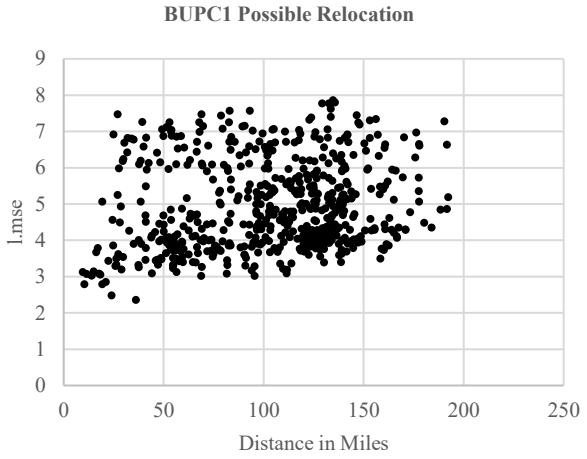


Figure 12: BUPC1 (Candidate Re-location) Distance versus MSE of SMART Mapping by Site

Another suspect site is E3738. For the location given by MADIS, we get M_L = max(0, 0, 0, 0, 0, 0) = 0. The candidate optimal location yields M_L = max(0.67, 0.33, 0.30, 0.38, 0.32, 0.36) = 0.67. See Figure 13 and Figure 14. This site also appears to be mis-located.

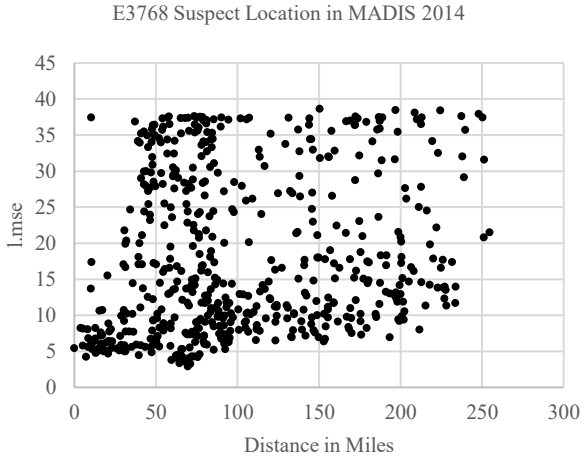


Figure 13: E3768 (Likely Incorrect Location) Distance versus MSE of SMART Mapping by Site

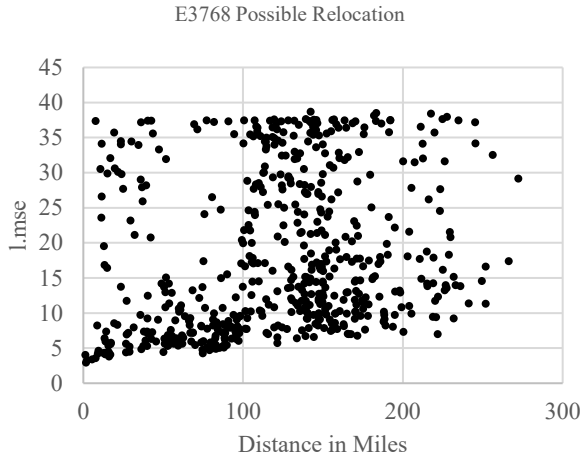


Figure 14: E3768 (Candidate Re-location) Distance versus MSE of SMART Mapping by Site

These are just several of the sites we identified in the 2014 data set that appear to mis-located and that have not subsequently been corrected by the provider. We suspect that the problem is a large one, greater than what is reflected by subsequent changes in the data, with some sites dramatically mis-located and others by lesser amounts. As demonstrated, MADIS did update locations of many sites between 2014 and 2019, so perhaps the problem is lesser now. It is hard to tell since we truly do not know with certainty which sites are mis-located.

We looked at the 2019 data and examined 625 sites meeting the same criteria as described for the 2014 data and identified further examples of what we suspect to be mis-located sites. Our analysis is not complete, but it does raise suspicion that the problem has not been alleviated. Figure 15 and Figure 16 show 2019 data for site PSWC1 and they appear to show that the site is mis-located. This is just one of multiple examples we identified using the logic presented above.

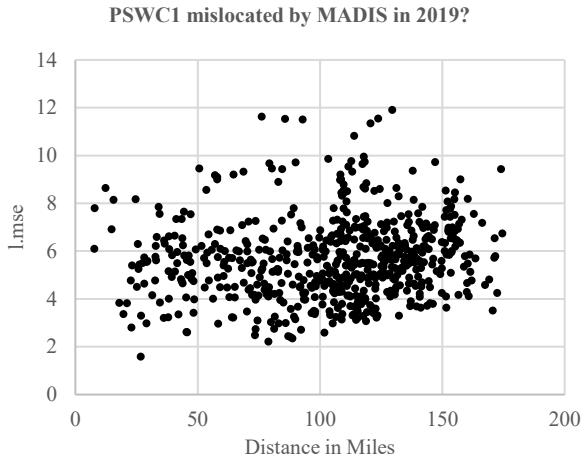


Figure 15: PSWC1 (2019 Location) Distance versus MSE of SMART Mapping by Site

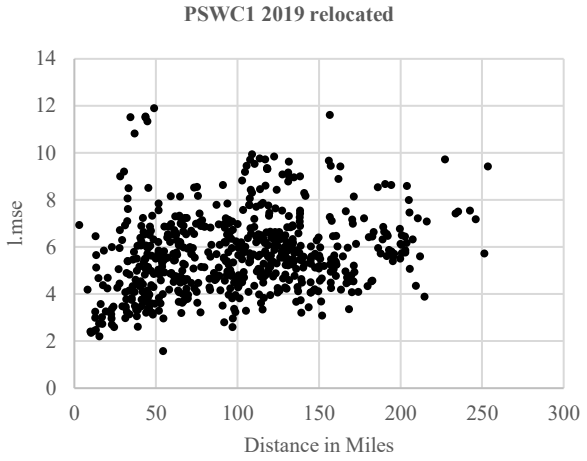


Figure 16: PSWC1 (Candidate Re-location) Distance versus MSE of SMART Mapping by Site

There is more work to do. The analysis above seems to work, but it could work better. Tightening the parameters used for the SMART mappings may help, and this could be done relative to individual sites. Iteration could be incorporated by removing the most suspect sites and recomputing. And, there would be value in removing individual data points identified as bad. All of this would help but would add to the complexity of the approach. Further investigation is merited.

VII. CONCLUSION

While our SMART method out-performed the other methods in nearly all instances in terms of accuracy of prediction of original data and classification of bad data, it was not our intent present it as the "best" method. Instead, we presented it as representative of the type of approach needed to overcome challenges of spatio-temporal data quality assessment.

It makes no assumption of isotropic covariance and does not require the determination of a specific covariance function. While it requires preprocessing time, it is suitable for near-real-time, online use. It accounts for disparate reporting times and frequency of reporting across sites. It not only helps to identify "bad data", but it also works well in the presence of bad data. It helps to identify and mitigate erroneous observations, "bad sites", and bad metadata. It uses multiple, robust methods to mitigate the impact of bad data on its estimates. Other methods, such as LSR and the various kriging approaches, could (and should) be modified in a similar manner to produce better, more robust results. Further, it is important to recognize the impact of bad data quality labels on evaluation. It is necessary to develop and use benchmark datasets with known, correct data quality labels.

A further advantage of our SMART method is that the SMART mappings provide a robust measure for comparing dissimilarity of sites. In turn, we showed how the SMART mappings could be used to identify mis-located sites. We

demonstrated that our approach works with known mis-located sites. We also demonstrated that there may be many mis-located sites for which the locations have not been corrected. Further work needs to be done in this area.

In general, we demonstrate that the quality assessment process must be an iterative process, with continual improvement and data incorporated. Figure 17 illustrates this process in general terms. A critical component in this process is evaluation.

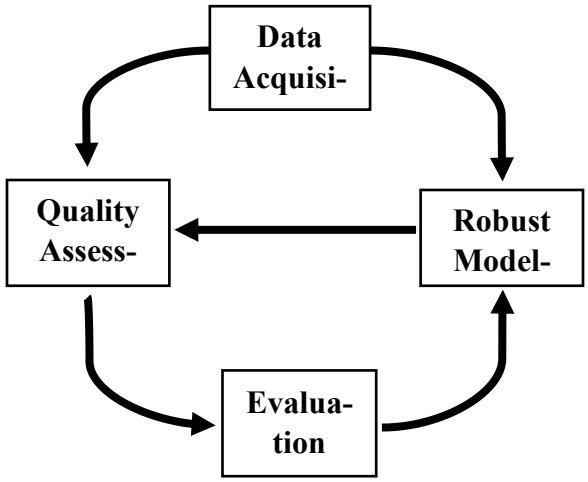


Figure 17: Iterative Data Quality Assessment Process

In this research, we investigated relatively simple situations and data sets involving ambient air temperature. We intend to expand our work to further examine other measures including wind and precipitation, as well as CCTV camera images. Departments of Transportation use CCTV camera images to verify road weather conditions reported by sensors. Yet, these images also suffer from poor data quality. Further research is needed to develop methods for detecting bad CCTV image data and for using CCTV image data to confirm sensor conditions and vice-versa. We intend to further develop benchmark datasets with known, good data quality labels.

REFERENCES

[1] D. E. Galarus and R. A. Angryk, "Challenges in Evaluating Methods for Detecting Spatio-Temporal Data Quality Issues in Weather Sensor Data," in *GEOProcessing 2019*.

[2] NOAA, "Meteorological Assimilation Data Ingest System (MADIS)." [Online]. Available: http://madis.noaa.gov/. [Accessed: 26-Dec-2015].

[3] NOAA, "MADIS Meteorological Surface Quality Control." [Online]. Available: https://madis.ncep.noaa.gov/madis_sfc_qc.shtml. [Accessed: 26-Dec-2015].

[4] U. of Utah, "MesoWest Data." [Online]. Available: http://mesowest.utah.edu/. [Accessed: 26-Dec-2015].

[5] U. of Utah, "MesoWest Data Variables." [Online]. Available: http://mesowest.utah.edu/cgi-bin/droman/variable_select.cgi. [Accessed: 26-Dec-2015].

[6] M. E. Splitt and J. D. Horel, "Use of multivariate linear regression for meteorological data analysis and quality assessment in complex terrain," in *Preprints, 10th Symp. on Meteorological Observations and Instrumentation, Phoenix, AZ, Amer. Meteor. Soc*, 1998, pp. 359–362.

[7] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[8] R. Nisbet, G. Miner, and J. Elder IV, *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[9] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[10] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education, Inc., 2006.

[11] C. C. Aggarwal, *Outlier Analysis*. Springer Publishing Company, Incorporated, 2013.

[12] S. Shekhar, C. T. Lu, and P. Zhang, "A unified approach to detecting spatial outliers," *Geoinformatica*, vol. 7, no. 2, pp. 139–166, 2003.

[13] A. Klein and W. Lehner, "Representing Data Quality in Sensor Data Streaming Environments," *J. Data Inf. Qual.*, vol. 1, no. 2, pp. 1–28, 2009.

[14] A. Klein and W. Lehner, "How to Optimize the Quality of Sensor Data Streams," *Proc. 2009 Fourth Int. Multi-Conference Comput. Glob. Inf. Technol. 00*, pp. 13–19, 2009.

[15] A. Klein, "Incorporating quality aspects in sensor data streams," *Proc. {ACM} first {Ph.D.} Work. {CIKM}*, pp. 77–84, 2007.

[16] A. Klein, H. H. Do, G. Hackenbroich, M. Karnstedt, and W. Lehner, "Representing data quality for streaming and static data," *Proc. - Int. Conf. Data Eng.*, pp. 3–10, 2007.

[17] A. Klein and G. Hackenbroich, "How to Screen a Data Stream." [Online]. Available: http://mitiq.mit.edu/ICIQ/Documents/IQ Conference 2009/Papers/3-A.pdf. [Accessed: 26-Dec-2015].

[18] S. L. Barnes, "A technique for maximizing details in numerical weather map analysis," *J. Appl. Meteorol.*, vol. 3, no. 4, pp. 396–409, 1964.

[19] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," *23rd ACM Natl. Conf.*, pp. 517–524, 1968.

[20] M.E. Splitt and J. Horel, "Use of Multivariate Linear Regression for Meteorological Data Analysis and Quality Assessment in Complex Terrain." [Online]. Available: http://mesowest.utah.edu/html/help/regress.html. [Accessed: 26-Dec-2015].

[21] U. of Utah, "MesoWest Quality Control Flags Help Page." [Online]. Available: http://mesowest.utah.edu/html/help/key.html. [Accessed: 26-Dec-2015].

[22] NOAA, "MADIS Quality Control." [Online]. Available: http://madis.noaa.gov/madis_qc.html. [Accessed: 26-Dec-2015].

[23] S. L. Belousov, L. S. Gandin, and S. A. Mashkovich, "Computer Processing of Current Meteorological Data, Translated from Russian to English by Atmospheric Environment Service," *Nurklik, Meteorol. Transl.*, no. 18, p. 227, 1972.

[24] D. Zimmerman, C. Pavlik, A. Ruggles, and M. P. Armstrong, "An experimental comparison of ordinary and universal

kriging and inverse distance weighting," *Math. Geol.*, vol. 31, no. 4, pp. 375–390, 1999.

[25] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Comput. Geosci.*, vol. 34, no. 9, pp. 1044–1055, 2008.

[26] T. G. Mueller, N. B. Pusuluri, K. K. Mathias, P. L. Cornelius, R. I. Barnhisel, and S. a. Shearer, "Map Quality for Ordinary Kriging and Inverse Distance Weighted Interpolation," *Soil Sci. Soc. Am. J.*, vol. 68, no. 6, p. 2042, 2004.

[27] D. E. Galarus, R. A. Angryk, and J. W. Sheppard, "Automated Weather Sensor Quality Control.," *FLAIRS Conf.*, pp. 388–393, 2012.

[28] D. E. Galarus and R. A. Angryk, "Mining robust neighborhoods for quality control of sensor data," *Proc. 4th ACM SIGSPATIAL Int. Work. GeoStreaming (IWGS '13)*, pp. 86–95, Nov. 2013.

[29] D. E. Galarus and R. A. Angryk, "A SMART Approach to Quality Assessment of Site-Based Spatio-Temporal Data," in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '16)*, 2016.

[30] D. E. Galarus and R. A. Angryk, "The SMART Approach to Comprehensive Quality Assessment of Site-Based Spatial-Temporal Data," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 2636–2645.

[31] D. E. Galarus and R. A. Angryk, "Beyond Accuracy - A SMART Approach to Site-Based Spatio-Temporal Data Quality Assessment," *(Accepted). Intell. Data Anal.*, vol. 22, no. 1, 2018.

[32] D. E. Galarus and R. A. Angryk, "Quality Control from the Perspective of the Real-Time Spatial-Temporal Data Aggregator and (re)Distributor," in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*, 2014, pp. 389–392.

[33] D. E. Galarus and R. A. Angryk, "Spatio-temporal quality control: implications and applications for data consumers and aggregators," *Open Geospatial Data, Softw. Stand.*, vol. 1, no. 1, p. 1, 2016.

[34] H. Xie, K. T. McDonnell, and H. Qin, "Surface reconstruction of noisy and defective data sets," in *Proceedings of the conference on Visualization'04*, 2004, pp. 259–266.

[35] L. Li, X. Zhou, M. Kalo, and R. Piltner, "Spatiotemporal interpolation methods for the application of estimating population exposure to fine particulate matter in the contiguous US and a Real-Time web application," *Int. J. Environ. Res. Public Health*, vol. 13, no. 8, p. 749, 2016.

[36] J. Grieser, "Interpolation of Global Monthly Rain Gauge Observations for Climate Change Analysis," *J. Appl. Meteorol. Climatol.*, vol. 54, no. 7, pp. 1449–1464, 2015.

[37] N. Cressie, "The origins of kriging," *Math. Geol.*, vol. 22, no. 3, pp. 239–252, 1990.

[38] H. Wackernagel, *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2013.

[39] M. S. Handcock and M. L. Stein, "A Bayesian analysis of kriging," *Technometrics*, vol. 35, no. 4, pp. 403–410, 1993.

[40] G. J. Hunter, A. K. Bregt, G. B. M. Heuvelink, S. De Bruin, and K. Virrantaus, "Spatial data quality: problems and prospects," in *Research trends in geographic information science*, Springer, 2009, pp. 101–121.

[41] E. H. Isaaks and R. M. Srivastava, *An introduction to applied geostatistics*. Oxford University Press, 1989.

[42] C. Huijbregts and G. Matheron, "Universal kriging (an optimal method for estimating and contouring in trend surface analysis)," in *Proceedings of Ninth International Symposium on Techniques for Decision-making in the Mineral Industry*, 1971.

[43] M. Galassi and Et-al, *GNU Scientific Library Reference Manual (3rd Ed.)*. Free Software Foundation.

[44] G. Bohling, "Introduction to Geostatistics and Variogram Analysis." [Online]. Available: http://people.ku.edu/~gbohling/cpe940/Variograms.pdf.

[45] P. J. Rousseeuw and K. Van Driessen, "Computing LTS regression for large data sets," *Data Min. Knowl. Discov.*, vol. 12, no. 1, pp. 29–45, 2006.

[46] R. F. Voss, "Random fractal forgeries," in *Fundamental algorithms for computer graphics*, Springer, 1985, pp. 805–835.

[47] J. Feder, *Fractals*. Springer Science & Business Media, 2013.

[48] M. F. Barnsley *et al.*, *The science of fractal images*. Springer Publishing Company, Incorporated, 2011.

[49] M. F. Goodchild and S. Gopal, *The accuracy of spatial databases*. CRC Press, 1989.

[50] D. E. Galarus, "Modeling stock market returns with local iterated function systems," 1995.

[51] T. C. Bailey and A. C. Gatrell, *Interactive spatial data analysis*, vol. 413. Longman Scientific & Technical Essex, 1995.

[52] USGS, "USGS Water Data for the Nation." [Online]. Available: https://waterdata.usgs.gov/nwis/. [Accessed: 28-Apr-2017].

eLIF: European Life Index Framework - An Analysis for the Case of European Union Countries

Ilie Cristian Dorobăț, Vlad Posea
Faculty of Computer Science
Politehnica University of Bucharest
Bucharest, Romania
Email: ilie.dorobat@stud.acs.upb.ro, vlad.posea@cs.pub.ro

Abstract—**With the continuous evolution of society, the analysis of the quality of life of the population has become an increasingly complex process, for which it is necessary to evaluate not only the factors that measure the financial power and the degree of economic development of the region, but also of those through which it can be appreciated the integration of individuals in society and of their implication within the well-functioning community. The importance of such an analysis is revealed from the implications of the insufficiency or even lack of measures to improve the standard of living has on members of society. Thus, as a result of the need of determining the living conditions, the implementation of the European Life Index Framework has been proposed. The Framework aims to automate the process of determining the quality of life of the population, data which the public authorities can use to easily determine the necessary steps for integrating disadvantaged people, reducing the poverty rate of the population, and improving quality of life. After analyzing the level of quality of life in European Union for the period 2007-2017, we have noticed that in the case of the former communist states, the quality of life standard is lower than that of the states which had a political trajectory outside the influence of the communist dictatorial regime. Also, due to the public policies mainly oriented towards citizens, the Nordic states have registered the highest values of the Quality of Life Index, surpassing even the countries of Continental Europe.**

Keywords-quality of life index; quality of life dimensions; open data quality; eLIF Framework.

I. INTRODUCTION

Measuring the population's standard of living is an important instrument in determining the degree of development of a region, just as a high level of quality of life reflects both the well-being of individuals as members of society and the well-being of society as a whole [1]. Moreover, a high degree of satisfaction of individuals regarding the level of quality of life has an important influence not only on the increase of the well-being of the population, but also on the economy, by increasing the productivity of the work and the correct remuneration of the personnel, raising both public and private capital investments increasing the number of jobs and improving working conditions, as well as encouraging a healthy lifestyle approach.

Although Gross Domestic Product (GDP) is the most used indicator of economic performance measurement [2][3], it cannot reflect by itself the population's standard of living; measuring living standards being a complex process that must take into account not only the economic but also the social side. Therefore, for the most appropriate calculation of the Quality of Life Index (QoLI), it had been included not only dimensions that include economic and financial indicators, but also dimensions that reflect the degree of security of citizens, the ability of medical units to provide specialized medical assistance, the degree of development of the educational system, the integration of the population in the field of work, the level of relation of the individuals as well as other indicators that have a significant influence on both the economic and social welfare.

The European Life Index Framework (eLIF) [4] is designed as a (semi) automatic QoLI calculation system, relieving specialists from identifying and applying complex sets of calculations. Thus, they have at their disposal a system whose task is to calculate the QoLI level from the perspective of eight objective dimensions and a subjective one, whilst the analysts only have to download the set of data regarding the outcome of the parliamentary elections provided by the International Institute for Democracy and Electoral Assistance (IDEA) and, of course, interpret the final result .

The present case of study considers the analysis of the level of quality of life of the population in the European Union, for each individual Member State, over a period of 11 years, starting with the year of Romania's accession to the European Union (2007), up to the last year for which statistical data are available for most of the analyzed indicators (2017). For a better understanding of the analysis context, we will make a brief presentation of the most important indicators of measurement of the level of development used over time, followed in Section III by a focus on the presentation of the nine economic and social dimensions used for the QoLI calculation.

Further, in Section IV, the intent is to present both the data sets used in the calculation of the standard of living of the population and the sources from which the data were obtained, and then in Section V the topic "quality assessment" will be approached. Thus, it will be performed an analysis of the data sets in order to identify the "quality

issues" and the means used to correct them. Section VI is intended for the presentation of the architecture of the eLIF framework, where we will detail the calculation formulas applied both for determining the QoLI value and for determining the values of the dimensions that fall within its composition. Next, Section VII is intended to present the actual analysis of the standard of living of the European Union population for the 11 years analyzed, and the last section is reserved for exposing of a series of findings regarding the approached subject.

II. STATE OF ART

The GDP is the most used indicator of economic performance measurement [2][3] in the intertemporal and interspatial comparison, expressing the balance of the total economy's production account as the sum of gross value added of the various institutional sectors and the various activity branched and taxes, from which are subtracted the subsidies on products, which are not allocated by the industry [5]. On a different note, as it can be observed in (1), the GDP is the economic indicator calculated from the perspective of the overall volume of consumption expenses, both from the governmental and the private area, of the governmental expenses, of the investments and of the trade balance.

$$GDP = \mathrm{C} + \mathrm{Gov} + \mathrm{I} + \mathrm{Export} - \mathrm{Import} \quad (1)$$

C = The households' expenses;
Gov = The central and local administration expenses;
I = The value of the investments;
Export = The sum of the expenses made by foreigners for goods and services produces in the country (exported);
Import = The value of the expenses of the residents for goods and services produces outside their country (imported).

From the financial point of view, the GDP reflects the economic development degree of an administrative-territorial unit for a given period of time, usually a semester or a year. Nevertheless, in comparative analysis, in its form, the GDP loses its accuracy because the ratio of the natural or legal persons which generates the GDP may enregister fluctuations even for the same analysed administrative-territorial unit. For this reason, its derivate, the GDP per capita, represents the standard instrument for this kind of analysis. Thus, the assessment of the GDP per capita may be interpreted as being not only an increase of the economic development level, but also an increase of the population's life quality level from various perspectives, as follows:

a) The increase of consumption may result both from the populations income increase, and from the applications of the governmental measures of stimulation of the consumption increase, increase which can influence the investments level and exports;
b) The increase of governmental expenses fuels, most of the times, the turnover level of the private area;
c) The increase of the level of the public and private investments may determine both the labour productivity assessment through re-technologization, through the increase of the employees' abilities, etc., and the assessment of the contractors' turnover;
d) The assessment of the trade balance may influence the life quality level through the increase of the availability of capital, capital which can be reinvested or used for the employees and/or shareholders fidelity.

Even though the GDP is an indicator often used in intertemporal and interspatial analysis regarding the degree of economic development, it cannot fully reflect the population's standard of living. Thus, starting from the fundamental needs, specialists have identified a wider range of factors which have an important influence on the standard of living [6], that is: i) financial stability; ii) health and safety; iii) interpersonal relationships; iv) the individual role in society; v) personal development.

Another metric for determining the development level of a certain region which has captured the attention of the politic decisions factors [7] has its origins in the year 1990, when United Nations Development Programme (UNDP) launched a new formula, by means of which the result is calculated by considering the factors which influence the richness of human life, and not the economy in which human beings live [8]. Thus, for determining the population's standard of living, the indicator suggested by UNDP, the Human Development Index (HDI) [9], uses the following three dimensions fundamental for the human development [1]:

a) the population's health status and longevity;
b) the level of knowledge the citizens have access to;
c) the access to resources necessary for a decent standard of life.

Even though for the modern man the financial part might represent a factor with a strong influence over the lifestyle adopted, in the HDI calculation, this factor does not have a greater significance than the other two. Therefore, as it can be observed in (2), the three factors which have a significant influence over the richness of human life are aggregated through the means of geographical average, a mathematic procedure which ensures the proportional distribution of the three factors.

$$HDI = \sqrt[n]{Health * Edu * Income} \quad (2)$$

Health = The population's health status and longevity;
Edu = The level of knowledge the citizens have access to;
Income = The level of financial resources to which the individual has access to for sustaining a decent standard of life;
n = The total number of indicators taken into account (three indicators).

The acceptance of the multidimensional nature of the factors which have a significant influence over the life standard which an individual adopts has led to the emergence of a new metric, World Health Organization Quality of Life (WHOQOL) [10]. This new method of determining the population's welfare, implemented by the World Health Organization (WHO) with the aim of identifying and protecting the vulnerable persons, takes into account more dimensions than HDI, that are:

a) physical domain;
b) psychological domain;
c) level of independence;
d) social relationships;
e) environment;
f) spirituality, religion, personal beliefs.

For developing the WHOQOL 15 cultural centers from different countries have participated, which had the role of applying the set of questionnaires realized for this purpose to a sample of 300 people complying to the following structure: i) half the interviewed persons are aged under 45 and the other half are aged over 45; ii) half the sample of people are male and the other half are female; iii) 250 are persons with a disease or impairment and 50 "healthy" respondents. Finally, in order to analyze the variation between different domain predictors for the criterion of quality of life, a regression analysis has been performed [11].

Even though HDI and WHOQOL are two metrics of a great importance to be taken into account in the analysis for determining the level of economic development and population's standard of life, the European Union's Statistical Office (Eurostat) has suggested that, in official reporting, the measurement of the population's welfare will be realized from the perspective of eight objective dimensions and one subjective dimension [12]. This set of dimensions, also known as "8+1 dimensions", being composed by the following:

a) material and living conditions;
b) productive or main activity;
c) health;
d) education;
e) leisure and social interactions;
f) economic and physical safety;
g) governance and basic rights;
h) natural and living environment;
i) overall experience of life.

As far as the sphere of ensuring the data quality is concerned, in literature [13][14][15], the following four main dimensions can be distinguished through which it can be ensured a level high as possible of the quality of data [1]: i) the data accuracy measures the degree of representativeness of the data stored in databases against the real world's elements which they represent; ii) the data consistency refers to the data's property of respecting the integrity constraints; iii) the information completeness measures the database's capacity of providing complete information to the user's query; iv) the data currency reflects the degree of the data's actualization.

III. Quality of Life Indicators

Even though GDP has been used for a long period as an intertemporal and interspatial comparison metric of the degree of economic development and the population's life standard [2][3], this indicator offers results strictly from the financial perspective. Therefore, considering the measurement of the life quality transcends the financial aspect, for the calculation of the life standard, scientists have tried to identify and to take into consideration all factors which have an important influence [16]. Thus, starting from the financial aspect and up to the citizens' own opinion, Eurostat groups these factors in 8 dimensions relative to the functional capacities which citizens must have for a decent life standard, and a subjective dimension relative to individuals personnel perspective on the personal achievement of life satisfaction and well-being [12].

A. Material and Living Conditions

Considering the complexity of life, the living standards are associated more like to the real income of the population and with the environment in which they live, rather than the GDP. Therefore, in order to measure these factors which influence the population's standard of living, Eurostat proposes the use of the Material and Living Conditions (MLC) dimension, through which the level of living is reflected, not only from a financial perspective, but also from the point of view of the living conditions.

If the financial aspect can be easily measured by reporting the purchase power of the population, through the determination of the median of incomes and through the identification of the inequality of income distribution (S80/S20 income quintile share ratio), determining the living conditions implies an analysis process of different factors which have a major influence of the individuals' social life. These factors reflect, on the one hand, the environment in which the analysed population lives, and on the other hand, the difficulty of satisfying the basic needs and of a decent living, as well as the individuals' capacity of sustaining the expenses necessary to enable them to have a decent living [1] (contracting mortgage loans, paying bills, purchasing long use goods, traveling inside and outside the frontiers, owning an automobile, etc.).

B. Productive or Main Activity

With the acceptance on wide scale of money as means of exchange, trading goods and services became a simpler process, and the individuals' attention was oriented towards developing and improving personal and professional skills. Therefore, in tandem with the society's evolution, each individual must allocate a significant part of his/her personal time to provision of labour to ensure the financial source

necessary both for sustaining everyday expenses and to engage in different social and professional activities.

Productive or Main Activity (PMA) is a separate dimension, built both from the perspective of the quantity and of the quality of the employment, which envisages the identification of the effects the professional life has on individuals. From the point of view of the quantity, the unemployment and the long-term unemployment rate are two factors which have a significant influence in determining the population's living standard because, as it has been related in the European Committee's Report [17], "people who become unemployed report lower life-evaluations, even after controlling for their lower income, and with little adaptation over time; unemployed people also report a higher prevalence of various negative affects (sadness, stress and pain) and lower levels of positive ones (joy). These subjective measures suggest that the costs of unemployment exceed the income-loss suffered by those who lose their jobs, reflecting the existence of non-pecuniary effects among the unemployed and of fears and anxieties generated by unemployment in the rest of society".

As far as the quality aspect is concerned, the PMA includes a series of entheogen indicators through which can be used to measure the benefits gained as a result of employment, the overqualification of the workforce, the equilibrium between professional and personal life (the number of working hours per week and the proportion of people working night shifts), the discrimination in the workplace, the safety at work. At the same time, besides the indicators which can be identified as being quantitative or qualitative, the PMA also includes two other factors found at the boundary between the two categories, involuntary temporary work and involuntary part-time employment.

C. Health

Health is a dimension which becomes more and more important with increasing age because the prevalence of the chronic diseases tends to increase as we age, by the increase of the life expectancy and the efficiency of the treatments against disease and conditions determine an increasingly stronger bound between the Health dimension and the determination of the population's living standards level [18]. On the other hand, this dimension has also economic prevalence, not only in establishing the budget for prevention and population treatment actions, but also from the human resources perspective, which, if it does not have the capacity necessary for employment, it becomes from a supplier of added value in a beneficiary of treatment services.

Being a complex dimension, more categories of factors are taken into consideration, beginning from the measurement of the healthy food consumption, up to determining the level of health infrastructure. Thus, embedded within this indicator, there are, in the one hand, the proportion of the population consuming daily fruits and vegetables, and on the other hand, the proportion of the population with unhealthy habits. Directly linked with these factors, there are both the life expectancy at birth and the health expectancy at birth, which measure the average number of years a new-born lives, respectively the average number of healthy years which a new-born lives, as well as the proportion of the population which is involved in physical activities and the effective healthy life, which measure the proportion of the population which considers to be in relatively good and very good health.

Offsetting the indicators which measure the hope of life and healthy life of the population, there are the indicators which measure the proportion of the population which has a long-standing illness or health problem, the proportion of population which cannot afford to support health analysis (including the dental ones), the incidence of the occupational accidents which need medical recovery for more than 4 days, and the proportion of overweight population. As long as the infrastructure is concerned, the number of hospital beds per 100,000 inhabitants and the proportion of medical personnel per 100,000 inhabitants represent veritable instruments to measure the capacity of the medical system to serve the population in the context of insuring the needed treatments.

D. Education

Education, as a dimension which describes the process of assimilation of knowledge and of improving the personal skills, represents the foundation of the human society, having, at the same time, a major impact upon the individuals' life quality [12]. Therefore, a solid level of education can favour the population in identifying and accessing some well-paid jobs, which contribute to the possibility of accessing high quality medical services and to the increase of living conditions. Furthermore, the risk of social exclusion and the poverty level can be diminished, and the degree of the population's implication in the public life, both as simple citizens and as political decision markers, may experience a favourable assessment.

Despite the importance which this dimension plays in the individuals' life, from a scientific point of view, the measurement of the population's educational level represents a complex process, and the mere reporting to the quantitative measure of years of schooling may not have the desired effect due to the fact that this indicator does not reflect the level of accumulated knowledge [19]. Thus, besides measuring the number of years which a student spent inside the education system, it is also necessary to analyse the factors which reflect the knowledge development and the cognitive skills.

E. Leisure and Social Interactions

If the time spent for the professional carrier development represent a sacrifice which each individual has to accept in order to beneficiate of a stable income source, the rest of the time is dedicated to household activities and recreative activities destined to improve the mental and physical health, to improve the self-esteem and self-confidence, to create social support and consolidating family bounds [20]. All these elements are found in the Leisure and Social

Interactions (LSI) dimension, dimension which has the role of determining the evolution of the self-esteem and the degree of participation of the individuals' in society with the help of two categories of indicators: i) indicators of measurement of the degree of the individuals' implication in society; ii) indicators referred to the personal bounding (family, friends, neighbours).

Thus, the degree of participation of the individuals in cultural and sportive activities; the proportion of the population which do not participate at these activities due to financial considerations or lack of infrastructure; the degree of participation of the individuals in the volunteer activities, are all indicators through which it can be determined the degree of the individuals' involvement in society. As far as the second category is concerned, which envisages the estimation of the support which individuals can receive at need, there are used as calculation elements both the proportion of persons which have relatives, friends and neighbours on which to rely for moral, material and financial support, and the proportion of persons which have at least a person with whom may discuss personal matters.

F. Safety

Safety is a state of stability both social and economic [1], which allows individuals to feel free of menaces and to concentrate on the personal and professional activities in which they are engaged. From a wide perspective, the Safety dimension envisages the measurement of the impact which safety risks to which the population is subject to on their welfare, being structured under the following criteria: i) economic safety; ii) physical safety.

As the International Committee of The Red Cross defines it [21], the Economic Safety is reflected through the individuals, households or community's capacity to cover with dignity the expenses generated by the satisfaction of the primary needs. In order to express these factors, in the Economic Safety calculation can be considered, on the one hand, the power of purchase of the retirees and the proportion from the GDP of the expenditure of social protection (administrative expenses only), which reflect the level of the income sources of the elderly persons and of the ones in exceptional situations (unemployed, persons with disabilities, families with low income etc.), and on the other hand, the proportion of the population incapable of coping with some unexpected financial expenses or in arrears.

Physical Security is the component of the Safety dimension through which it is evaluated the level of protection of the individuals in front of crimes which may affect the physical and mental integrity of the victims or, through which the victims may be illegally dispossessed of personal goods. Such crimes which can be taken into account in the calculation of the physical security level are assaults, kidnaps, sexual violence, robberies and thefts, traffic and consumption of heavy drugs etc.

G. Governance and Basic Rights

Governance and Basic Rights (GBR) incorporates a series of factors which influence the level of the population life standard from the perspective of governance, regulation and guarantee of equal rights between the community's individuals regardless of their health status, financial state, political, religious or cultural orientation. Therefore, for calculating this dimension, can be considered indicators such as employment gender gap and gender pay gap, which measure the existing differences on labour market between generations; the degree of trust the population has in the political and legal system and in Police; parliamentary voter turnout etc.

H. Natural and Living Environment

Pollution represent one of the world's biggest problems due to the fact that its effects are increasingly felt, so that just for the year 2015 it has been estimated that 9 million cases of premature death (16% of all deaths worldwide) have been caused by the effects of pollution [22]. From air pollution cause by consumption of fossil fuel in industries and transportation area, to the pollution of the groundwater as a result of the toxic waste storage and to the acoustic pollution recorded in crowded cities, they all case harmful effects on the surrounding environment and on the population life standards.

Thus, in a world ever more polluted, in which the effects of global worming are felt with an ever more increased intensity, the peoples' need and acknowledgement to protect the surrounding environment become a task ever more important both for the governmental institutions as well as for the nongovernmental organizations. For this, starting from determining the level of chemical and acoustic pollution up to determining the proportion of population which have access at least to one drinkable water source, the Natural and Living Environment (ENV) dimension represents a valuable instrument for determining the level of the quality of the surrounding environment in which individuals live and undergo their activities; results which the political decision factors may use to identity and applicate solutions aimed at conserving and improving the quality of the environment.

I. Overall Experience of Life

The statistical data are collected and prepared to serve as information and analysis resource both for political factors implied in the planification and evaluation of political decisions, as well as for private organizations and population which have the right to be informed regarding the evolution of the society they live in. Nevertheless, because no objective indicators can perfectly measure the described concept, in order to determine the population's living standard, the subjective wellbeing measurement gains a particular importance due to the fact that through this dimension can be realised an overall image of the society groups which perceive the living standards as good or bad [23].

IV. OPEN DATA SOURCES

In a world found in a continuous evolution, in which the speed of the information dissemination is at the distance of a click, and the population becomes more and more consciousness regarding the power of knowledge, ensuring the free access to data becomes a task ever more important for the worldwide institutions. Even though the term open data may imply that it defines those data which are not restricted in use, in their reuse and distribution, some suppliers might have different perspective regarding what openness represents [24]. Thus, even though the access to data is free of charge, the actions of use, reuse, reworking, redistribution and reselling might be limited or restricted through the terms and conditions imposed by the data suppliers [1].

In the governmental area, the Open Government Data initiatives started to fall into place, so that ever more states supply data with free access for users. Nevertheless, even though each state has its own rules and priorities of data publication, the existence of some aggregators as Eurostat facilitates the access of interested persons to sets of public data. Thus, Eurostat make available both a user-friendly interface, as well as an API Server [25], through which the process of obtaining the sets of data can be automated.

For obtaining the data relative to the 8+1 analysed dimensions, the programmatic interface made available by Erostat has been used, so that the analysis can be easily be extended for any desired period of time. At the same time, due to the fact that the data relative to the parliamentary voter turnout is not available in the Eurostat statistics, these data have been obtained from reliable suppliers as the International Institute for Democracy and Electoral Assistance (IDEA) [26]. The name of all of these data sets can be seen in Table I, where we summarized them in order to make available to the audience the used indicators.

TABLE I. DATA SOURCES.

Dimension Name	Dimension Indicator Name
Material and Living Condition (10 JSON files)	Dwelling Issues Rate End Meet Inability Rate High Income Rate Income Quintile Rate Material Deprivation Rate Over Occupied Rate Poverty Risk Purchasing Rate Under Occupied Rate Low Work Intensity Rate
Productive or Main Activity (9 JSON files)	Average Work Hours Employment Rate Involuntary Part-Time Rate Long Term Unemployment Rate Researchers per Ten Thousand Inhabitants Temporary Employment Rate Unemployment Rate Working Nights Rate
Health (12 JSON files)	Fruits and Vegetables Consumption Rate Health Personnel per Ten Thousand Inhabitants Healthy Life Rate Healthy Life Years - Female Healthy Life Years - Male Hospital Beds per Ten Thousand Inhabitants Life Expectancy Long Health Issues Rate Obese Population Rate Smokers Rate Unmet Dental Rate Unmet Medical Rate Work Accidents per Thousand Inhabitants
Education (9 JSON files)	Digital Skills Rate Early Education Rate Education Rate Excluded Rate School Dropout Rate Students to Teachers Rate Training Rate Zero Foreign Language Rate
Leisure and Social Interactions (6 JSON files)	Asking Rate Discussion Rate Getting Together Rate Non-participation Rate to Cultural Activities or Sports Events due to important reasons Participation Rate to Cultural Activities or Sports Events Participation Rate to Voluntary Activities
Economic and Physical Safety (6 JSON files)	Crime Rate Offences per Thousand Inhabitants Pension Power Social Protection Power Unexpected Financial Expenses Rate Nonpayment Rate
Governance and Basic Rights (4 JSON files, 1 CSV file)	Active Citizenship Rate Employment Rate Gender Pay Gap Parliamentary Elections Participation Rate [a] Population Trust Rate
Natural and Living Environment (2 JSON files)	Noise Pollution Rate Pollution Rate
Overall Experience of Life (1 JSON file)	High Life Satisfaction Rate
Auxiliary Dimensions (1 JSON file)	Population on 1 January

[a] Data set downloaded from the portal of IDEA

V. OPEN DATA QUALITY ASSESSMENT

In the literature, the concept of data quality is referred to as an indicator by which data utility can be measured from the perspective of data consumers [27][28], a broad term used to describe this concept being “fitness for use” [27][29][30]. At the core of the process of measuring the quality of the data are the data producers and the data custodians, whose role is to generate the data, respectively storage them, to ensure the

maintenance and the security of the data, so as the data customers can use them in the provided form or after applying some processes of data aggregation, and data integration [28].

Starting from the significance of the open data concept, we can deduce that open data quality is part of the concept of data quality that concerns the data with free access for use, regardless of the type of license under which they are provided [1]. Regarding the practical way of measuring the quality of open data, the most widespread dimensions in the literature are accuracy, completeness, consistency and timeliness [13][28][31].

A. Data Currency Issue

Data currency or timeliness is an indicator of measuring the quality of the data used to determine the degree of the currency of data in relation to the specific activity for which they are used [15]. As in the case of the API provided by the National Institute of Statistics of Romania, the API provided by Eurostat presents the same deficiency: the update date is available for the data set as entities and not for records from data sets [1]. Therefore, although we may have a reference regarding when to update the data sets, we cannot identify whether they have been modified as a result of adding new records or as a result of updating existing records.

B. Data Inconsistency

The inconsistency of the data can be defined as the lack of data consistency, meaning that state of the data in which the format and value are not in accordance with the chosen data model [32] or which have discontinuities [1]. In the case of the current analysis, the inconsistency of the data is materialized both by the discontinuity in time of the data sets, and by their different format, Eurostat offering a flexible API Server that returns the data sets using the JSON-stat standard [33], while IDEA offers the possibility to export data on the results of parliamentary elections in "xls" format. Thus, as can be seen in Table II, the share of missing data is slightly over 31%, a result determined primarily by the presence of indicators for which data is available only for one year.

For example, in the case of the Overall Experience of Life dimension, which has a single indicator - the High Satisfaction Rate, the total number of values expected to be present in a data set without discontinuities is 308 ($28\ countries * 11\ years$). However, because values are available only for the year 2013, the share of missing data is very high, reaching about 91% $\left(\frac{28\ countries * 10\ years}{28\ countries * 11\ years} * 100\right)$ of the total data. For other data sets, the periods with discontinuities may be shorter, but the lack of data even for a single year for a state prevents us from determining the QoLI value for that particular state.

Therefore, to correct the data discontinuity, the present study proposes that the value for missing years be supplemented with the value of the previous year, and if for any previous year there is no value, the value of the following year will be assigned. The advantage of using this approach instead of calculating the average of the series [1] is that the value thus calculated is closer to "truth", that is, the average of the series can be much higher or much lower compared to the fluctuation of the values from one year to the following.

TABLE II. ENTRIES STATISTICS.

Dimension Name	Available Values	Expected Values [a]	Missing Data (%)
Material and Living Condition	3,352	3,388	1.06
Productive or Main Activity	2,449	2,464	0.61
Health	2,951	4,004	26.30
Education	1,985	2,464	19.44
Leisure and Social Interactions	448	4,312	89.61
Economic and Physical Safety	2,786	3,080	9.55
Governance and Basic Rights	1,088	2,464	55.84
Natural and Living Environment	610	616	0.97
Overall Experience of Life	28	308	90.91
Auxiliary Dimensions	532	560	5.18
TOTAL	**16,229**	**23,660**	**31.41**

[a] The total number of entries that should exist for the data set to be complete

As for the format of the data sets, if in the case of those provided by Eurostat a standard format is used which allows quickly querying and processing of data, the same cannot be said about the data set provided by IDEA. The latter is saved in "xls" format, having all data stored in string type columns. Moreover, the percentage data is presented as a numerical value followed by the percentage sign "%", which requires that, before converting to numerical format, a cleaning operation of the values is performed so that they can be used in mathematical operations.

C. Lack of Data

Determining the level of quality of life of the population involves the study of economic, financial and social phenomena and processes in which individuals are engaged and which influences their lives. To this end, identifying official sources and data sets which reflect the factors of influence is the first step in conducting such an analysis. Unfortunately, although Eurostat provides a wide range of data sets from different fields of activity and areas of interest, some data sets are not complete, making it impossible to perform a comparative analysis in time and space.

By incomplete data sets are defined the ones for which data from at least one country is missing for the entire analyzed period of time (2007-2017). Thus, considering the principle listed in the previous subsection regarding the completion, in Table III we find name of data sets that were completely excluded from the QoLI calculation.

VI. THE FRAMEWORK ARCHITECTURE

Determining the level of quality of life of the population is a complex process which, due to the impact it can have in determining and applying social and economic policies, has attracted the attention of more and more researchers. Having as a starting point a series of indicators that, over time, have gained visibility in this direction, the eLIF [4] has been developed to provide specialists with a complex (semi) automatic solution for calculating Quality of Life Index from the perspective of 8 objective dimensions and a subjective one. Thus, the persons interested in carrying out analyzes regarding the state of living of the population are relieved from identifying and implementing the QoLI calculation formulas; the only tasks being to download the data set regarding the outcome of the parliamentary elections provided by IDEA and to interpret the obtained results.

Regarding the design architecture of the eLIF framework, as can be seen from Figure 1, the architecture is conceived in four steps: i) data preprocessing; ii) calculating the values of the QoLI dimensions; iii) preliminary analysis of the result; iv) presentation of the result.

Data processing, as a preliminary step in calculating QoLI dimensions, it begins by identifying official sources and data sets that reflect the main factors that influences the quality of life. Based on both the sources and the data sets, as well as the analysis carried out in the previous sections, the following step is to trace the 8+1 dimensions that have a major influence on the quality of life of the population, and then the attention is focused on identifying and implementing the data structure on which the subsequent calculations will be performed.

TABLE III. THE LACK OF DATA BY COUNTRIES FOR ALL OF THE ANALYZED PERIOD.

Index Name	Missing Data by Countries
The share of population exposed to air pollution	Malta
The share of population connected to public water supply	Italy Latvia Slovenia United Kingdom
The share of population who consumes alcohol daily	France
The share of population who did aerobic and muscle-strengthening exercises	Belgium Netherlands
The share of population who declared that are over-qualified employees	Denmark Ireland Netherland
The share of population having neither a bath, nor a shower, nor indoor flushing toilet in their household	Sweden

Before performing the actual calculations, an important step for any analysis is data preparation, that is to say processing them to meet the own needs. This process involves both performing various corrections on the data (cleaning the data, converting the numerical values into percentage values, consolidating several indicators into one etc.), as well as filling in the missing data with a replacement value using the approach presented in the previous section.

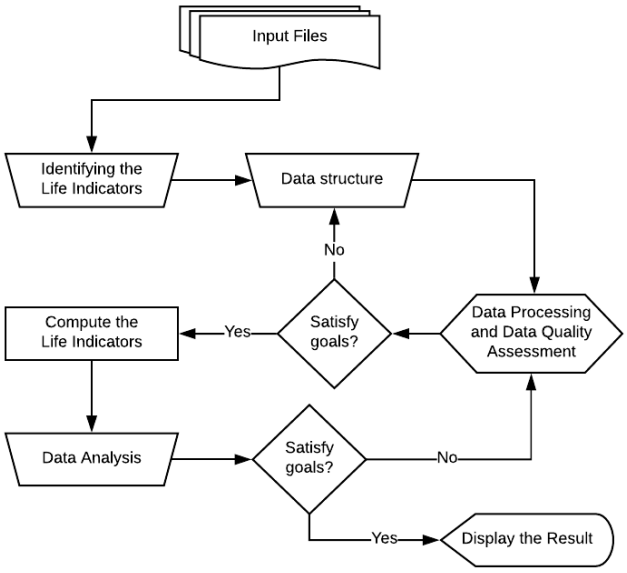


Figure 1. The framework architecture.

However, if during this process it is found that the data structure no longer corresponds to the practical reality, the procedure returns to the (re)design stage of the data structure, and later (re)processing. Subsequently, after processing the data extracted from the official sources, we proceed to *calculating the values of the 8+1 QoLI dimensions*, a stage in which the formulas for calculating the dimensions reach the final form and are applied to the processed data in order to generate the analysis data.

The third stage, *preliminary data analysis*, is an intermediate stage for presenting the final result, with the purpose of identifying potential errors and discrepancies in the process of calculating the indicators [1]. For example, End Meet Inability Rate, Material Deprivation Rate, Unemployment Rate, Unmet Dental / Medical Rate, School Dropout Rate etc. are indicators which have a negative influence on the calculation of the dimensions they belong to, which requires in the final formula the use of synthetic indicators that determine the weight of the population that is not affected (R - Reversed Rate), using the calculation formula set out in Figure 2.

```
Function getReversedRate (value)
    return 100 - value;
```

Figure 2. Function for getting the Reversed Rate.

Thus, after correcting the values of the indicators that have a significant influence in determining the level of quality of life, the flow of operations returns to the processing of the data to recalculate the final values of the QoLI dimensions. Finally, the last stage, *displaying of results*, allows users to extract the final result broken down by countries and years both for the generic QoLI indicator and for the dimensions

that compose it. The actual calculation of these values is performed by applying the logarithmic function to the product of the indicators / dimensions related to the measured metric as in (3).

$$indicator = \ln(p) \quad (3)$$

p = The product of dimensions/indicators.

In the case of QoLI, the value is calculated as a result of applying the logarithmic function previously presented over the product of all the dimensions that compose it as in (4).

$$p = MLC * PMA * Health * Education * LSI * Safety * GBR * Environment * Overall\ Exp \quad (4)$$

MLC = Material and Living Conditions dimension;
PMA = Productive or Main Activity dimension;
Health = Health dimension;
Education = Education dimension;
LSI = Leisure and Social Interactions dimension;
Safety = Safety dimension;
Environment = Natural and Living Environment dimension;
Overall Exp = Overall Experience of Life dimension.

The use of the logarithmic function for the calculation of the final result is required by the asymmetrical character of the dimensions that belong to QoLI, respectively of the indicators that belong to the composition of the QoLI dimensions, so that it can be avoided the case when a low value indicator has to be compensated by another high value indicator [1]. After a closer analysis, in contrast to the paper [1], in which, for the calculation of the indicators, it has been chosen to extract the root of the order n, we considered that the logarithmic function is a truer instrument because, as J. Martin Bland and Douglas G Altman acknowledge [34], data transformation through logarithmic function offers the most interpretable results even after applying the anti-log function to cancel the logarithmic calculation result. Thus, using the logarithmic function to calculate the QoLI result we assure that a 1% change in one dimension will have the same impact as the 1% change in any other dimension.

A. Material and Living Conditions

MLC is one of the most complex dimensions taken into account when determining the level of quality of life of the population, because, besides determining the financial conditions of the population, through this dimension, the level of living conditions of the population can also be determined. In order to determine the MLC value, the calculation of the product of all the indicators related to this dimension is considered as in (5), followed by the application of the logarithmic function over this product.

$$p = R(DWI\ Rate) * R(EMI\ Rate) * High\ Income\ Rate * R(Quintile\ Rate) * R(Deprivation\ Rate) * R(Over\ O\ Rate) * R(Poverty\ Risk\ Rate) * PPS\ Rate * Under\ O\ Rate * R(WI\ Rate) \quad (5)$$

DWI Rate = Proportion of the population living in dwelling with a leaking roof, damp walls, floors or foundation, or rot in window frames of floor (%);
EMI Rate = Proportion of the households making ends meet with difficulty and great difficulty (%);
High Income Rate = Proportion of the population having income of 130% of median income or more (%);
Quintile Rate = Inequality of income distribution (S80/S20 income quintile share ratio) (%);
Deprivation Rate = Severe material deprivation rate (%);
Over O Rate = Overcrowding rate (%);
Poverty Risk Rate = At-risk-of-poverty rate (%);
PPS Rate = Purchasing Power Standard as percent of the European Union countries (%);
Under O Rate = Share of people living in under-occupied dwellings (%);
WI Rate = Share of people up to 59 years living in households with very low work intensity (%).

B. Productive or Main Activity

As a dimension that measures the quality of life from the perspective of the professional side of individuals, Productive or Main Activity includes a series of indicators regarding the worked hours, the type of the accepted work contracts and the share of the unemployed population. Thus, for the calculation of this dimension, the logarithmic formula will be applied to the product of the indicators found in the PMA composition as presented in (5), with the mention that the Average Work Hours (AWH) indicator will be processed to determine the weight of the number of hours worked at every 12 hours per day over a week according to (6).

$$p = C(AWH) * Emp\ Rate * R(Inv\ Part\ Time) * R(L\ T\ Unemp\ Rate) * R(Nights\ Rate) * Researchers\ Rate * R(Temp\ Emp\ Rate) * R(Unemp\ Rate) \quad (5)$$

$$C(AWH) = (12\ hours * 7\ days) - AWH \quad (6)$$

AWH = Average number of usual weekly hours of work in main job worked by full-time employed persons aged 15 years or over (number of hours);
Emp Rate = Percentage of employed people aged from 15 to 64 years (%);
Inv Part Time = Involuntary part-time employment as percentage of the total part-time employment (%);

L T Unemp Rate = Percentage of long-term unemployed people aged from 15 to 74 years (%);
Nights Rate = Percentage of the total employment aged from 15 to 64 years who are working at nights (%);
Researchers Rate = Full-time equivalent researchers per ten thousand inhabitants;
Temp Emp Rate = Percentage of total employment who are working based on temporary contracts (%);
Unemp Rate = Percentage of labour force aged 15-74 years who are unemployed (%).

C. Health

Health is one of the dimensions that has always been a reference for the scientific world in the process of determining the standard of living of the population, carrying out a wide range of analysis of the implication of the level of health on the respondents [35][36][37]. As can be seen in (7), for the current analysis we considered a series of factors that have an important influence on the population both from the perspective of the health system's ability to provide specialized healthcare as well as the perspective of the population to adopt a healthy lifestyle.

Thus, there can be identified both indicators that measure the lifestyle of the population (Fruits and Vegetables Consumption Rate, Obese Population Rate, Smokers Rate) as well as indicators that show the access of the population to health services (Health Personnel, Hospital Beds) and that estimate the expectations regarding the level of health of the respondents (Healthy Life Rate, Life Expectancy, Work Accidents, etc.).

$$p = R(Obese\ Rate) * FV\ Rate * Personnel\ Ratio * HL\ Rate * HLY\ Female * HLY\ Male * H\ Beds * Life\ Expectancy * R(LTH\ Issues\ Rate) * R(Smokers\ Rate) * R(UDS) * R(UMS) * R(WA\ Rate) \quad (7)$$

Obese Rate = Percentage of people who are obese (%);
FV Rate = Share of population that consumes fruits and vegetable daily (%);
Personnel Rate = Health personnel (medical doctors; nurses and midwives; dentists; pharmacists; physiotherapists) per hundred thousand inhabitants;
HL Rate = Share of people aged 16 years and over who are self-perceiving very good or good health;
HLY Female = Female health expectancy at birth;
HLY Male = Male health expectancy at birth;
H Beds = Hospital beds per hundred thousand inhabitants;
Life Expectancy = The number of remaining years a person is expected to live at birth or at a certain age;
LTH Issues Rate = Share of people aged 16 years or over having a long-standing illness or health problem;
Smokers Rate = Share of people who smokes cigarettes daily;
UDS = Share of people who self-reported unmet needs for dental examination;
UMS = Share of people who self-reported unmet needs for medical examination;
WA Rate = Work accidents per ten thousand inhabitants.

D. Education

Even if the short-term impact within the Education dimension is not no visible, as a primary factor in the development of both society as a whole and of individuals in private, this dimension has a significant influence on establishing the standard of quality of life. Thus, in the long term, through a solid education system, which takes into account both group and individual needs, the influence of education can be reflected both by the development of the individual character of the population and by the good training of professionals; whom can then be integrated more easily into the field of work. As in the calculation formulas of the other dimensions, Education is calculated by applying the logarithmic function presented in (3) to the product of the indicators specific to this dimension as presented in (8).

$$p = Digital\ Skills * Early\ Edu\ Rate * R(Excluded\ Rate) * R(School\ Dropout\ Rate) * Students\ Rate * Pupils\ Rate * Training\ Rate * NKFL\ Rate \quad (8)$$

Digital Skills = The share of people (aged from 16 to 74 years) who have basic or above basic overall digital skills (%);
Early Edu Rate = The share of pupils aged between 4 years old and the starting age of compulsory education who are participating in early childhood education (%);
Excluded Rate = The share of people (aged from 18 to 24 years) neither in employment nor in education and training (%);
School Dropout Rate = The share of people (from 18 to 24 years) who leave education and training early (%);
Students Rate = The share of people (aged from 15 to 64 years) who are participating in tertiary education level (%);
Pupils Rate = Ratio of pupils to teachers for primary and secondary education (%);
Training Rate = The share of people (aged from 25 to 64 years) who are participating in education and training in the last 4 weeks (%);
NKFL Rate = The share of people (from 25 to 64 years) who don't know any foreign language (self-reported).

E. Leisure and Social Interactions

As the leisure time is related to social activities, in calculating this dimension are taken into consideration both the indicators that reflect the moral support that the population can receive from close persons, and the indicators by which the social activity of the individuals is reflected.

Thus, the general calculation formula presented in (3) applies over the product of all these indicators as in (9).

$$p = Asking\ Rate * Discussion\ Rate * Getting\ Together\ Rate * R(Non\ Participation\ Rate) * Social\ Activities\ Rate * Voluntary\ Activities\ Rate \quad (9)$$

Asking Rate = The share of people (aged 16 years or over) who have someone to ask for help (moral, material or financial) from family, relatives, friends or neighbors (%);

Discussion Rate = The share of people (aged 16 years or over) who have someone to discuss personal matters (%);

Getting Together Rate = The share of people (aged 16 years or over) getting together with friends at least once a week (%);

Non Participation Rate = The share of people (aged 16 years or over) who are not involved in cultural activities or sports events during the previous 12 months due to financial reasons or due to a lack of facilities (%);

Social Activities Rate = The share of people (aged 16 years or over) who are involved in any cultural or sport activities in the last 12 months (%);

Voluntary Activities Rate = The share of people (aged 16 years or over) who are involved in formal or informal voluntary activities (%).

F. Safety

The Safety dimension is of particular importance because it can determine the degree of safety of the population, both physically and financially. Therefore, for the calculation of this dimension, the product of the indicators that compose it will be used as in (10), with the mention that the indicators Pension Power and Social Protection Power will be corrected by dividing them by 100 as in (11).

$$p = R(Crime\ Rate) * C(Pension\ Power) * C(Social\ Protection\ Power) * R(Unexpected\ Rate) * R(Non\ Payment\ Rate) * R(Offences\ Rate) \quad (10)$$

$$C(value) = \frac{value}{100} \quad (11)$$

Crime Rate = The share of the population who perceived there was crime, violence or vandalism in the area where they live (%);

Pension Power = The average pension (Purchasing Power Standard per inhabitant);

Social Protection Power = Social protection expenditure (Purchasing Power Standard per inhabitant);

Unexpected Rate = The share of the population unable to face unexpected financial expenses (%);

Non Payment Rate = The share of the population in arrears on mortgage or rent, utility bills or hire purchase (%);

Offences Rate = Recorded offences (assault, robbery, sexual offences, theft, unlawful offences) per thousand inhabitants.

G. Natural and Living Environment

Natural and Living Environment is the only one dimension that, for calculating the Quality of Life Index, takes into account indicators through which the state of the surrounding environment is reflected. Although green space per capita is an important indicator for measuring the mental health of individuals [38][39], in the absence of an official data set, two other indicators that measure the quality of the environment will be taken into consideration, namely: noise pollution, and respectively pollution. In order to determine the quality index of the environment, the aggregation of these indicators is performed in a similar way to the aggregation of the other dimensions' indicators, by applying the logarithmic function over the product of the related indicators as in (12).

$$p = R(Noise\ Pollution\ Rate) * R(Pollution\ Rate) \quad (12)$$

Noise Pollution Rate = Share of population reporting noise from neighbours or from the street (%);

Pollution Rate = Share of population exposed to Pollution, grime or other environmental problems (%).

H. Overall Experience of Life

Unlike the other eight statistical dimensions of measuring the quality of life from the perspective of the objective functional capacities of individuals, the Overall Experience is the only dimension that takes into account people's choices, priorities and values [12]. Thus, the calculation of this dimension is done by means of a single indicator that measures the proportion of the population that experiences a high quality of life level, as in (13).

$$Overal\ Exp = \ln(High\ Satisfaction\ Rate) \quad (13)$$

High Satisfaction Rate = Share of population rating their overall life satisfaction as high (%).

VII. DATA USAGE

Estimating the level of quality of life is a complex process that involves monitoring not only of the economic, financial and environmental indicators through which the economic development and financial power of the population is reflected, but also the social indicators which reflect the degree of interrelationship of individuals in society. The importance of estimating this indicator stands in the very dimensions that come within its composition, being useful both to the government decision makers who have the legal levers for combating poverty and raising the standard of living of the population, as well as the other actors in the

society interested in following the annual evolution of the degree of economic and social development of the European Union Member States [1].

The present analysis aims to determine the level of quality of life of the population of European Union, starting with 2007, the year of Romania's accession to the European Union, until 2017, the last year for which statistical data are available for most of the analyzed indicators. Thus, we envisage the production of statistics that provide the interested parties with data on both top Member States that record a considerable advance and those with a lower level of quality of life in comparison with the other Member States.

Given that the final result of the study materializes into a comparative analysis of the quality of life for the 28 European Union Member States, for a period of 11 years (2007-2017), the final result of the data processing will be presented both in table form, through Table IV and in visual form, through Figure 3 and Figure 4. As it can be seen, the result of the calculation for determining the quality of life of the population can be divided into two sections, depending on the political ideology on which the states were governed before the 1990s. Thus, in the former communist states (Bulgaria, Czech Republic, Cyprus, Croatia, Estonia, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia, Hungary) as well as in four other states outside the sphere of influence of the communist regime (Greece, Malta , Portugal, Spain) a low level of quality of life of the population can be identified; most of the states that did not have a significant influence of the communist doctrine, registering a high level of quality of life of the population throughout the analyzed period of time.

One of the pillars of a modern society that has a significant influence on both the personal as well as on the professional life of individuals, is education, because, through continuous learning and improvement of their abilities, the members of society can more easily develop and benefit from interpersonal relationships and they can achieve better results in the field in which they operate. Unfortunately, although, theoretically, the early integration of children in the education system should enable them to develop their personal capacities and abilities, in practice, the quality of the education system has a determining influence. In this regard we can see that for the analysed period, although in Bulgaria and Romania the share of pupils aged between 4 years old and the starting age of compulsory education does not register values lower than 83%, in Finland, country where one of the best education systems in the world is found [40][41][42], this indicator starts below 70%, exceeding the 80% threshold only starting with 2012. At the same time, the participation rate of the adult population (adults between the ages of 25 and 64) in training courses reaches, in the Scandinavian countries, even over 30%, whereas, in the vast majority of the former communist states, this indicator does not even reach the 10% limit; in Bulgaria and Romania are registered the lowest values in the whole European Union, with values below the 3% threshold.

As in any economy, the quality of the education system is directly reflected through the level of purchasing power of individuals, more precisely, through the level of income in relation to their own needs; the population with a higher level of qualification can register a higher value of the remuneration of the work performed and, implicitly, the value of the future retirement pension will be higher. Thus, regarding the financial security of the population, we can see that, in the countries of the former communist bloc, both the purchasing power of the population and the general capacity of individuals to deal with unexpected expenses, register lower values than the countries in which the influence of the communist regime was not so great. Similar trends are also identified in other indicators that measure the financial stability and living conditions of the population such as inequality of income distribution (S80/S20 income quintile share ratio), severe material deprivation rate, share of people living in over-occupied dwellings, etc.

As regards health, a particular situation is encountered in the case of Bulgaria and Romania, in the meaning that, although in these two countries the level of life expectancy at birth indicator has some of the lowest values in the whole European Union unlike the countries outside the former communist bloc that registered the highest values, regarding the share of people (aged 16 years or over) having a long-standing illness or health problem, the situation is completely different, in the meaning that in Bulgaria and Romania have been registered the lowest share of the population with long-standing illness or health problem, being followed just a few places away by Denmark, while Finland is the country of the European Union where the highest proportion of such cases has been reported. Therefore, although the lowest life expectancy registered in the European Union is in Bulgaria and Romania, the resident population lives most of their lives without serious health problems and without incurable or very difficult to treat diseases, while, in Finland, one of the most developed countries, the population often faces such problems throughout their lives.

However, it should be noted that although the two countries at the end of the QoLI ranking have the lowest share of people suffering from serious illnesses, at the same time have the highest share of population who, for financial reasons or related to the distance at which the medical units are located, cannot benefit from the specialized treatment. This phenomenon is prevalent in the former communist countries that had a different trajectory of economic and social development compared to the western countries.

Neither with regard to the active involvement of population in the life of society, the countries of the former communist bloc do not register exemplary values, since in most of these states the level of involvement is below 17%. In Romania and Bulgaria have being registered the lowest rates, 3.2%, respectively 5.2%, while in developed countries the share of the population involved in volunteering activities can reach just over 30%, and in Scandinavian countries, even up to almost 40%. A similar statistic is also reflected in the

share of the population engaged in cultural or sporting activities. From this perspective, Romania and Bulgaria are also on the last two places, with values below the 30% threshold (Bulgaria rising to 31.9% in 2015), while most of the other former communist states registering values between 40% and 70%. On the other side, the countries where the influence of the communist doctrine did not have such an important impact, the share of the population participating in various cultural and sporting activities exceeds 70%, the top being occupied by the Scandinavian countries with values of over 80% registered over the whole time period taken into consideration.

Analysing the trends that the indicators that make up the QoLI ranking registered during 2007-2017, we can notice that the European Union is divided into three important groups divided according to the economic and social evolution of each country. The first group is made up of the majority of the former communist states that, in the process of transition from a dictatorial regime to an open regime, have faced various specific challenges such as guaranteeing and protecting the rights, destructuring of oppressive institutions, liberalizing the market, strengthening relations with The West, attracting foreign capital, etc. Thus, as can be seen in Figure 4, the last places are occupied by Bulgaria and Romania, which, for the entire analyzed period, maintained their last places (28th and respectively 27th), followed by Latvia, which until 2011 ranked 26th. After 2012, against the background of applying the limiting measures of the negative effects of the public debt crisis (freezing wages and eliminating bonuses in the public sector; raising and introducing new taxes; reducing salaries and pensions, etc.) [43], Greece was to record a significant decrease in the quality of life, bringing the country to the 26th place, while Latvia was to rise of one place, occupying the 25th at the end of the analyzed period.

The second largest group is mainly made up of Western countries, which have always had a trajectory oriented towards freedom. Although they are not in the top of the ranking, the experience of a lasting democracy serves as a support for a stronger economy, thus, the success of these countries may be determined by several public policies implemented over time, of which we list the following [44]:

- reducing variations and conditions for granting different levels of social protection such as unemployment benefit, social assistance etc.;
- orientation towards programs for induction within the field of work of young people, women and workers with limited abilities;
- emphasizing on policies for balancing professional and personal life.

Finally, the podium is occupied by the Nordic countries (Denmark, Finland, Sweden, Norway - which is not a member of the European Union) which have always had a very high quality of life compared to other European countries due to their ability to quickly adapt to new political, economic and social changes. The Nordic model of development can be characterized from the perspective of the following three key features [45]:

- stateness: the Nordic political classes understand better than the other European countries that the state does not have to be an oppressive apparatus;
- universalism: financial services and benefits are not only aimed at the needy, but also extend to the middle class;
- equality: equal opportunities are one of the values that the Nordics emphasize upon, thus, the Nordic countries have a high level of gender equality.

In the end, based on the evolution of the QoLI indicator of the states at the end of the ranking presented in Figure 4, we can see that they are registering an upward trend, with the exception of Greece, which, due to the public debt crisis, was forced to apply a series of austerity measures so as to avoid a potential bankruptcy of the whole country. Continuing at this rate, with the help of both the expertise of the developed countries and with the free sources of funding that the European Union offers to the Member States in order to cover the development gaps between regions (European funds), the less developed countries have the chance to quickly recover some of this gap. However, even if from 2013 the growth rate has become more and more rapid, for a solid development, these states must rethink the long-term development strategy, taking Poland as example, a country which has managed to become an important industrial center in Central and Eastern Europe, and Finland which made every effort to restructure the education system starting from the Swedish model.

On the other side of the ranking, the struggle is no longer carried out strictly in the direction of improving the quality of life of the population, but, rather, toward identifying and offering new opportunities and toward implementing measures to protect the people in difficulty. Thus, as can be seen from Figure 3, the evolution of the developed countries is slower, sometimes with a negative trend, with the exception of the states that are at the base of the ranking of the most developed countries. From this perspective, developed countries need to focus on improving their own systems, on importing and implementing solutions that work in other states (e.g. the Swedish model of education system implemented in Finland etc.) and on supporting disadvantaged countries because, in a globalized economy, the mutual development allows the fruition of the relations between partners.

Thus, having identified the areas where the standard of quality of life is low, political drivers can take the necessary measures to reduce the disparities between high-growth countries and those with low-growth. At the same time, non-governmental factors have at their disposal a set of data that allows them to monitor the activity of the decision-makers regarding the reduction of the gaps registered between the European Union Member States.

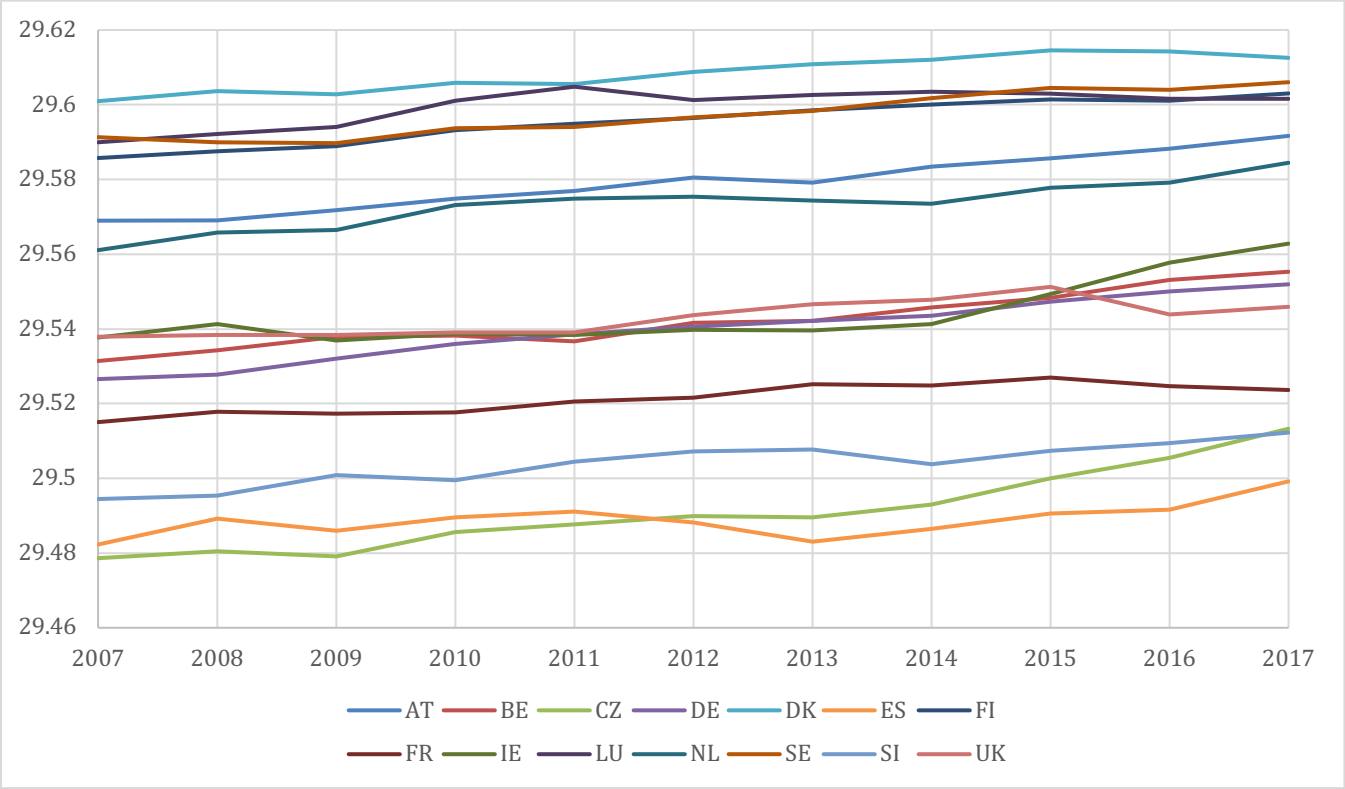


Figure 3. The heist levels of QoLI by years and countries.

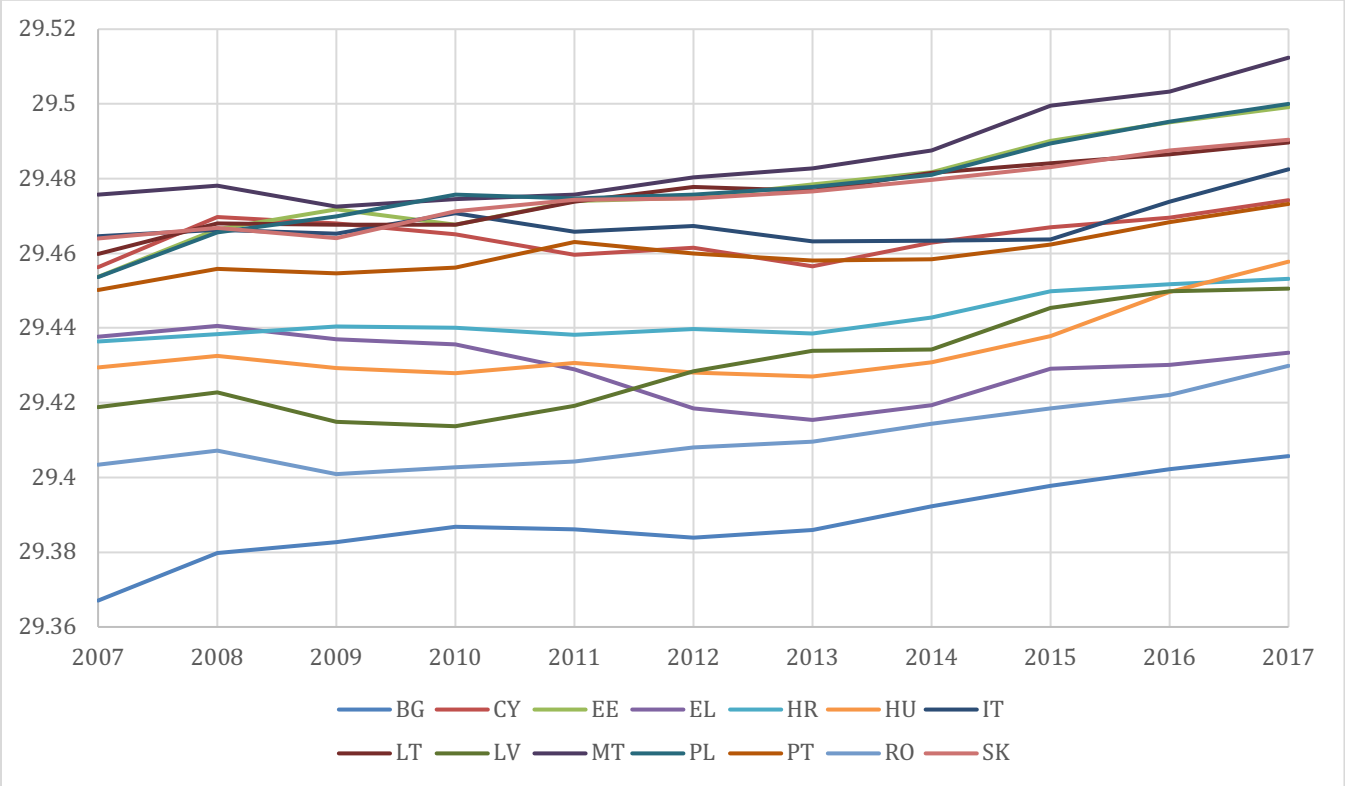


Figure 4. The lowest levels of QoLI by years and countries.

TABLE IV. THE QUALITY OF LIFE INDEX BY YEARS AND COUNTRIES.

Country / Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
[AT] Austria	29.5690	29.5690	29.5718	29.5749	29.5769	29.5804	29.5792	29.5835	29.5856	29.5882	29.5917
[BE] Belgium	29.5314	29.5342	29.5379	29.5382	29.5366	29.5416	29.5422	29.5458	29.5483	29.5531	29.5553
[BG] Bulgaria	29.3671	29.3797	29.3826	29.3869	29.3861	29.3839	29.3860	29.3923	29.3978	29.4022	29.4057
[CY] Cyprus	29.4563	29.4698	29.4680	29.4651	29.4597	29.4616	29.4566	29.4629	29.4670	29.4695	29.4742
[CZ] Czechia	29.4786	29.4805	29.4791	29.4857	29.4876	29.4899	29.4896	29.4929	29.5001	29.5055	29.5133
[DE] Germany	29.5266	29.5278	29.5321	29.5360	29.5385	29.5407	29.5421	29.5436	29.5474	29.5500	29.5519
[DK] Denmark	29.6009	29.6037	29.6028	29.6059	29.6056	29.6087	29.6109	29.6120	29.6145	29.6142	29.6125
[EE] Estonia	29.4537	29.4665	29.4719	29.4677	29.4739	29.4748	29.4784	29.4818	29.4902	29.4950	29.4991
[EL] Greece	29.4376	29.4406	29.4369	29.4357	29.4290	29.4186	29.4154	29.4194	29.4291	29.4301	29.4335
[ES] Spain	29.4823	29.4891	29.4859	29.4895	29.4911	29.4881	29.4830	29.4864	29.4906	29.4916	29.4992
[FI] Finland	29.5857	29.5875	29.5889	29.5931	29.5949	29.5965	29.5985	29.6000	29.6014	29.6011	29.6030
[FR] France	29.5150	29.5178	29.5173	29.5177	29.5206	29.5216	29.5252	29.5249	29.5270	29.5247	29.5236
[HR] Croatia	29.4364	29.4384	29.4404	29.4401	29.4381	29.4398	29.4386	29.4428	29.4498	29.4517	29.4532
[HU] Hungary	29.4295	29.4326	29.4293	29.4279	29.4306	29.4281	29.4270	29.4309	29.4378	29.4497	29.4578
[IE] Ireland	29.5376	29.5412	29.5369	29.5387	29.5384	29.5398	29.5395	29.5413	29.5494	29.5577	29.5628
[IT] Italy	29.4646	29.4663	29.4653	29.4708	29.4659	29.4673	29.4632	29.4634	29.4638	29.4738	29.4825
[LT] Lithuania	29.4598	29.4680	29.4677	29.4676	29.4738	29.4778	29.4768	29.4815	29.4841	29.4865	29.4897
[LU] Luxembourg	29.5899	29.5922	29.5941	29.6010	29.6048	29.6013	29.6026	29.6036	29.6030	29.6016	29.6016
[LV] Latvia	29.4189	29.4227	29.4149	29.4137	29.4192	29.4284	29.4339	29.4342	29.4453	29.4498	29.4506
[MT] Malta	29.4756	29.4782	29.4725	29.4745	29.4757	29.4803	29.4827	29.4876	29.4996	29.5034	29.5124
[NL] Netherlands	29.5611	29.5657	29.5664	29.5732	29.5748	29.5753	29.5743	29.5736	29.5778	29.5792	29.5845
[PL] Poland	29.4536	29.4656	29.4700	29.4758	29.4746	29.4758	29.4777	29.4810	29.4894	29.4953	29.5000
[PT] Portugal	29.4502	29.4558	29.4546	29.4562	29.4630	29.4599	29.4580	29.4585	29.4624	29.4684	29.4732
[RO] Romania	29.4034	29.4072	29.4009	29.4028	29.4043	29.4080	29.4096	29.4144	29.4185	29.4221	29.4299
[SE] Sweden	29.5914	29.5899	29.5897	29.5937	29.5940	29.5966	29.5983	29.6017	29.6045	29.6040	29.6060
[SI] Slovenia	29.4945	29.4953	29.5008	29.4994	29.5045	29.5072	29.5078	29.5037	29.5074	29.5094	29.5122
[SK] Slovakia	29.4640	29.4667	29.4641	29.4712	29.4743	29.4748	29.4766	29.4796	29.4830	29.4875	29.4904
[UK] United Kingdom	29.5379	29.5383	29.5384	29.5391	29.5391	29.5437	29.5465	29.5479	29.5512	29.5438	29.5459

VIII. CONCLUSION

The analysis of the standard of living of the population is a complex process that should not be limited only to determining the degree of economic development of the analyzed area but should be extended also to the social side by determining the degree of satisfaction that the analyzed group shows in society. Therefore, in addition to the economic and financial spheres that reflect the degree of economic development and the financial situation, when calculating the Quality of Life Index, analysts must also consider social indicators that reflect the level of contentment of individuals, the level of employee training and the degree of absorption in the field of work, the level of security that the state offers to the members within its society, the confidence of individuals amongst their peers and within the state institutions, etc.

By analyzing the data sets presented above, it can be noticed that the eastern part of the European Union, together with several states in the northern area, presents the lowest QoLI values; all these states having a common denominator: having been governed by a communist dictatorial regime until the 1990s. Unlike the countries of Continental Europe and Scandinavia, the lack of a modern vision, and the

transition from the communist era to the democratic one, can be a determining factor in terms of slower development. However, it is important to note that, although the former communist states had an additional impediment in the path to a harmonious development, for the analyzed period of time, all these states have a general upward trend of the QoLI level.

Therefore, the consolidation of the dimensions from both the economic and financial spheres as well as from the social one, in a single complex indicator such as QoLI, allows to carry out complex analysis regarding the level of development of some regions and the degree of population satisfaction. Thus, with the eLIF framework [4], through which the values of the quality of life index can be calculated for a given period, political factors have the possibility to identify the disadvantaged states that must be supported in order to achieve one of the objectives of the European Union, that of strengthening the economic, social and territorial cohesion and solidarity between Member States.

Also, this framework allows both non-governmental organizations to supervise the involvement of the government sphere in applying measures to increase the standard of living of the population, as well as to economic operators to identify areas with potential for development. Another important feature of the framework is its adaptability, which can be easily extended to any level of administrative detail by including in the analyzed list the name of the administrative unit targeted, whether it is a city, region, country or other form of administration, provided that the administrative unit is of the same type (it would not be feasible to compare the QoLI values recorded in counties with those of the regions or the values of any other types of different administrative units).

Regarding the further development direction, three main objectives will be considered: i) automatic integration of the results of the parliamentary elections data set, so as to fully automate the calculation process; ii) identification of alternative reliable data sources for completing the data sets excluded from the analysis (presented in Table III); iii) integrating both the QoLI result and the result of the dimensions that compose it into an information system similar to Visit Romanian Museums [46] touristic information system.

REFERENCES

[1] I. C. Dorobăț, O. Rinciog, G. C. Muraru and V. Posea, "Quality of Life Index Analysis for the Case of Romanian Regions", Proceedings of The Thirteenth International Conference on Digital Society and eGovernments (ICDS 2019), IARIA, Feb. 24-28, 2019, Athens, Greece, pp. 37-44, ISSN: 2308-3956, ISBN: 978-1-61208-685-9;

[2] P. Bartelmus. "Beyond GDP: New approaches to applied statistics", The Review of Income and Wealth, vol. 33 (4), Dec. 1987, pp. 347-358, doi:10.1111/j.1475-4991.1987.tb00679.x;

[3] V. Berenger and A. Verdier-Chouchane, "Multidimensional measures of well-being: Standard of living and quality of life across countries", World Development, vol. 35 (7), Jul. 2007, pp. 1259-1276, doi:10.1016/j.worlddev.2006.10.011;

[4] European Life Index Framework. [Online]. Available from https://github.com/iliedorobat/QoLI-Framework [retrieved: August, 2019];

[5] National Institute of Statistics of Romania. Tempo Online - GDP definition. [Online]. Available from http://statistici.insse.ro:8077/tempo-online/ [retrieved: August, 2019];

[6] J. C. Flanagan, "A research approach to improving our quality of life", American Psychologist, vol. 33 (2), pp. 138-147, Feb. 1978, doi:10.1037/0003-066X.33.2.138;

[7] E. Neumayer, "The Human Development Index and sustainability - a constructive proposal", Ecological Economics, vol. 39, pp. 101-114, Oct. 2001, doi:10.1016/S0921-8009(01)00201-4;

[8] United Nations Development Programme, Human Development Reports. About Human Development. [Online]. Available from http://hdr.undp.org/en/humandev [retrieved: August, 2019];

[9] United Nations Development Programme, Human Development Reports. Human Development Index. [Online]. Available from http://hdr.undp.org/en/content/human-development-index-hdi [retrieved: August, 2019];

[10] The WHOQL Group, "The World Health Organization Quality of Life assessment (WHOQL): position paper from the World Health Organization", Social Science & Medicine, vol. 41, pp. 1403-1409, Nov. 1995, doi:10.1016/0277-9536(95)00112-K;

[11] M. Power, M. Bullinger and A. Harper, "The World Health Organization WHOQOL-100: Tests of the Universality of Quality of Life in 15 Different Cultural Groups Worldwide", Health Psychology, vol. 18(5), pp. 494-505, Sep. 1999, doi:10.1037/0278-6133.18.5.495;

[12] Eurostat. Quality of life indicators. [Online]. Available from https://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators [retrieved: August, 2019];

[13] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations", Communications of the ACM, Nov. 1996, pp. 86-95, doi:10.1145/240455.240479;

[14] W. Fan, "Data quality: From theory to practice", ACM SIGMOD Record, Sep. 2015, pp. 7-18, doi:10.1145/2854006.2854008;

[15] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment", Communications of the ACM - Supporting community and building social capital, Apr. 2002, pp. 211-218, doi:10.1145/505248.506010;

[16] E. Diener and E. Suh, "Meaasuring quality of life: economic, social and subjective indicators", Social Indicators Research, vol. 40, pp. 189-216, Jan. 1997, doi:10.1023/A:1006859511756;

[17] J. E. Stiglitz, A. Sen and J. P. Fitoussi, "Report by the Commission on the Measurement of Economic Performance and Social Progress". [Online]. Available from https://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report/7bac2480-4658-439f-b022-e6542ebf714e [retrieved: August, 2019];

[18] A. Steptoe, A. Deaton and A. A. Stone, "Subjective wellbeing, health, and ageing", The Lancet, vol. 385(9968), pp. 640-648, Feb. 2015, doi:10.1016/S0140-6736(13)61489-0;

[19] E. A. Hanushek and L. Wößmann, The Role of Education Quality for Economic Growth. World Bank Policy Research Working Paper no. 4122. [Online]. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=960379 [retrieved: August, 2019];

[20] J. Kim, N. Yamada, J. Heo, A. Han, "Health benefits of serious involvement in leisure activities among older Korean adults", International Journal of Qualitative Studies on Health and Well-being, vol. 9(1), Jul. 2014, doi: 10.3402/qhw.v9.24616;

[21] International Commitee of the Red Cross. Economic security definition. [Online]. Available from https://www.icrc.org/en/what-we-do/ensuring-economic-security [retrieved: August, 2019];

[22] P. J. Landrigan, R. Fuller, N. J. R. Acosta, O. Adeyi, R. Arnold, N. Basu et al., "The Lancet Commission on pollution and health", The Lancet, vol. 391(10119), pp. 462-512, Feb. 2018, doi:10.1016/S0140-6736(17)32345-0;

[23] Organization for Economic Co-operation and Development, OECD Guidelines on Measuring Subjective Well-being. [Online]. Available from https://read.oecd-ilibrary.org/economics/oecd-guidelines-on-measuring-subjective-well-being/concept-and-validity_9789264191655-5-en#page1 [retrieved: August, 2019];

[24] Kitchin R., "The Data Revolution: Big Data, Open Data, data infrastructures and their consequences", SAGE Publications Ltd, pp. 48-67, 2014, ISBN: 978 1 4462 8748 4;

[25] Eurostat, The Rest Request. [Online]. Available from https://ec.europa.eu/eurostat/web/json-and-unicode-web-services/getting-started/rest-request [retrieved: August, 2019];

[26] International Institute for Democracy and Electoral Assitance, Voter Turnout Database. [Online]. Available from https://www.idea.int/data-tools/data/voter-turnout [retrieved: August, 2019];

[27] R. Y. Wang and D. M. Strong, "Beyound accuracy: what data quality means to data consumers", Journal of Mangement Information Systems, vol. 12 (4), Spring 1996, pp. 5-33, doi:10.1080/07421222.1996.11518099;

[28] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data quality in context", Communications of the ACM, May 1997, pp. 103-110, doi:10.1145/253769.253804;

[29] J. M. Juran, "Juran on Leadership for Quality: An Executive Handbook". Free Press, New York, 1989, ISBN 0029166829 9780029166826;

[30] C. W. Fishera and B. R. Kingma, "Criticality of data quality as exemplified in two disasters"; Information & Management, vol. 39(2), pp. 109 - 116, Dec. 2001, doi:10.1016/S0378-7206(01)00083-0;

[31] D. P. Ballau and H. R. Pazer, "Modeling data and process quality in multi-input information systems", Management Science, vol. 31, pp. 150-162, Feb. 1985, doi:10.1287/mnsc.31.2.150;

[32] B. Behkamal, M. Kahani, E. Bagheri, and Z. Jemeric, "A metrics-driven approach for quality assessment of linked open data", Journal of Theoretical and Applied Electronic Commerce Research, vol. 9, pp. 64-79, May 2014, doi:10.4067/S0718-18762014000200006;

[33] JSON-stat. [Online]. Available from https://json-stat.org/ [retrieved: August, 2019];

[34] J. M. Bland and D. G. Altman, "The Use Of Transformation When Comparing Two Means", British Medical Journal, vol. 312(7039), pp. 1153, May 1996, doi:10.1136/bmj.312.7039.1153;

[35] G. W. Torrance, "Utility approach to measuring health-related quality of life", Journal of chronic diseases, vol. 40(6), pp. 593-600, 1987, doi:10.1016/0021-9681(87)90019-1;

[36] E. Ware Jr. and B. Gandek, "Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project", Journal of clinical epidemiology, vol. 51(11), pp. 903-912, 1998, doi:10.1016/S0895-4356(98)00081-X;

[37] G. R. Norman, J. A. Sloan and K. W. Wyrwich, "Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation", Medical care, vol. 41(5), pp. 582-592, 2003, doi:10.1097/01.MLR.0000062554.74615.4C;

[38] S. de Vries, R. Verheij, P. Gronewegen and P. Spreeuwenberg., "Natural environments - Healthy environments? An exploratory analysis of the relationship between greenspace and health", Environment and Planning A: Economy and Space, vol. 35(10), pp. 1717-1731, 2003, doi:10.1068/a35111;

[39] P. Groenwegen, A. Van Den Berg, S. De Vries and R. Verheij, "Vitamin G: Effects of green space on health, well-being, and social safety", BMC Public Health, vol. 6(149), pp. 1-9, 2006, doi:10.1186/1471-2458-6-149;

[40] H. Morgan, "Review of Research: The Education System in Finland: A Success Story Other Countries Can Emulate", Childhood Education, vol. 90(6), pp. 453-457, 2014, doi:10.1080/00094056.2014.983013;

[41] P. Sahlberg, "The professional educator: lessons from Finland", American Educator, vol. 35(2), pp. 34-38, 2011, ISSN-0148-432X;

[42] P. Sahlberg, "A Model Lesson: Finland Shows Us What Equal Opportunity Looks Like", American Educator, vol. 36(1), pp. 20-27, 2012, ISSN-0148-432X;

[43] M. Matsaganis, "The welfare state and the crisis: the case of Greece", Journal of European Social Policy, vol. 21(5), pp. 501-512, Dec. 2011, doi:10.1177/0958928711418858;

[44] A. C. Hemerijck, "When Changing Welfare States and the Eurocrisis Meet", Sociologica. Italian Journal of Sociology (e-journal), vol 1(1), 2012, doi:10.2383/36887;

[45] A.W. Pedersen and S. Kuhnle, "The Nordic welfare state model", In: The Nordic models in political science: Challenged, but still viable?. Fagbokforl, pp. 249-272, 2017, ISBN 978-82-450-2175-2;

[46] O. Rinciog, I. C. Dorobăț and V. Posea, "Route Suggestion for Visiting Museums Using Semantic Data", eLearning & Software for Education, vol. 3, pp. 48-55, 2017, doi:10.12753/2066-026X-17-180.

Less-Known Tourist Attraction Analysis Using Clustering Geo-tagged Photographs via X-means

Jhih-Yu Lin
Graduate School of Systems Design
Tokyo Metropolitan University
Tokyo, Japan
lin-jhihyu@ed.tmu.ac.jp

Shu-Mei Wen
Department of Statistics and
Information Science in Applied
Statistics
Fu Jen Catholic University
Taipei, Taiwan
126531@mail.fju.edu.tw

Masaharu Hirota
Department of Information Science
Faculty of Informatics
Okayama University of Science
Okayama, Japan
hirota@mis.ous.ac.jp

Tetsuya Araki
Graduate School of Science and Technology
Gunma University
Gunma, Japan
tetsuya.araki@gunma-u.ac.jp

Hiroshi Ishikawa
Graduate School of Systems Design
Tokyo Metropolitan University
Tokyo, Japan
ishikawa-hiroshi@tmu.ac.jp

***Abstract*—Today, travelers can readily travel around the world using convenient transportation. Not only are opportunities to go abroad for sightseeing is increasing, but tourism industries of every country are developing indirectly. Moreover, many travelers obtain the latest tourist information from the internet for their journeys. However, most information specifically relates to popular tourist attractions, leading to crowds flocking there, which make tourists feel uncomfortable. Contrary to existing studies, which specifically emphasize analyses of popular tourist attractions, we are striving to disperse crowds from popular tourist attractions and provide more spots for travelers to choose by discovering less-known tourist attractions. This study therefore specifically examines discovery of less-known Japanese tourist attractions under the assumption that these spots exist in unfamiliar cities of tourists. According to results of analyzing geo-tagged photographs on Flickr, we use the X-means algorithm to group Japanese cities into different clusters. X-means is an extension of K-means that improved the shortcomings of K-means and which greatly reduced the probability of being trapped into a local optimum. Furthermore, these clusters were used to survey unfamiliar clusters to Japanese and Taiwanese people. Thereby, we can eliminate spots that are in familiar clusters. We propose a formula for ranking tourist attractions that lets travelers choose these spots easily. Results of verification experiments demonstrated that some less-known tourist attractions appeal to Taiwanese and Japanese. Additionally, we examined some factors that might affect respondents as they decide whether a spot is attractive to them or not.**

Keywords–Flickr; geo-tagged photograph; less-known tourist attractions；X-means

I. INTRODUCTION

In this era of the internet and smartphones, most people can readily share and record their tourist experiences on social networking services (SNSs) such as Facebook and Flickr. Numerous studies have analyzed user records of tours on SNS to elucidate user hobbies and preferences. In doing so, one can discover popular tourist attractions and recommend some tour plans for users according to their preferences [2]–[5]. Using SNSs, one can immediately obtain the newest status of friends, particularly using well-known functions related to check-in and "geo-tagged" photographs, which are useful when one wants to share a location with friends.

Aside from geolocation, diverse information is available from different SNS people users. That information includes important and useful data for research. For instance, Hausmann et al. [6] pointed out that social media contents might provide a swift and cost-efficient substitute for traditional surveys. Liu et al. [7] proposed an approach for the discovery of areas of interest (AOIs) by analyzing geo-tagged photographs and check-in information to suggest popular scenic locations and popular spots among travelers. Another study with similar aims to those of the present study used SNS users' information and geo-tagged photographs to suggest obscure sightseeing locations [8].

Most tourists receive sightseeing information through travel websites. However, almost all of these websites present well-known tourist attractions. Consequently, although the attractions are crowded and congested, visitors will be guided there. Our preliminary investigation revealed that most tourists do not like crowded spots that make them feel uncomfortable.

Many earlier studies have specifically addressed analyses of popular tourism attractions or AOIs while neglecting other unnoticed places. Our goal for this study is to improve several aspects through dispersal of crowds from more popular tourist attractions because (1) crowded popular tourist attractions make visitors feel uncomfortable, (2) foreign visitors are too numerous at popular tourist attractions, raising crime rates there, and (3) tourism to regions other than popular regions should be supported.

To accomplish our aim, we analyzed scenic geo-tagged photographs taken in Japan obtained from Flickr. After identifying some worthwhile and less-known tourist attractions, we examined them based on scenic photographs to assess their tourism value in terms of human landscapes, ecotourism, and natural landscapes. This study specifically examines natural landscapes: we used scenic photographs to appeal to travelers with natural landscapes. This study

therefore has a clearly defined research scope. Results can present more tourist attraction options for tourists and can reduce crowding at well-known tourist attractions.

Our earlier study [1] showed that over half of Taiwanese and Japanese respondents liked well-known tourist attractions and liked less-known tourist attractions. Also, questionnaire results indicated that income has little connection to travel frequency. Nevertheless, our earlier study has one point of possible improvement. Because of the influence of outliers, the grouping method used for the earlier study classified more than 70% of data into the same cluster, producing a drastically uneven data distribution. This study uses the X-means algorithm to ameliorate this shortcoming. Furthermore, we revised our earlier formula based on current questionnaire results.

The remainder of the paper is organized as follows: Section II introduces related work. Section III is an overview of the method. Section IV explains the scenic photograph evaluation method. In Section V, we present less-known tourist attraction estimation and explain our questionnaire results. Section VI presents survey questionnaire improvements and present conclusions and future work.

II. Related work

This section presents discussion of some studies related to our research, including benefits and risks of international tourism, POI and AOI, cluster analysis.

A. Benefits and Risks of International Tourism

Recently, tourism has become a development emphasis for many countries because international tourism can not only bring huge revenues; it can also have positive effects on increased long-run economic growth. Several reports have described that international tourism can bring benefits by promoting foreign exchange revenues, spurring investment in new infrastructure, stimulating other economic industries indirectly, and generating employment [9]–[14]. Moreover, Algieri et al. [15] reported that determinants of competitive advantages in tourism are important for both economically advanced and developing economies. Those determinants can help policy makers to design better strategies to strengthen activities exhibiting potential, improve performance, and enhance international competitive advantage, terms of trade, and economic growth.

Although the number of tourists continues to increase and bring huge revenues for tourism-related industries, benefits from tourism are accompanied by latent crises. Kakamu et al. [16] discovered that when the numbers of foreign visitors and the police force increase, the crime rate also increases. The rising crime rates can be expected to reduce willingness to visit and thereby tourism income [17].

B. POI and AOI

Points of interest (POIs) differ from areas of interest (AOIs). A POI is a particular spot that someone might find useful or interesting. They can be landmarks, sightseeing spots or commercial institutions of all types such as restaurants, hospitals, and supermarkets. Furthermore, POIs shown on a digital map must include some information such as name, type, longitude, and latitude. Based on data types and the discovery procedure, the approaches developed for POI are divided into two types. The first type is top-down: discovery of POI from an existing POI repository or database, such as check-in data or yellow pages that are frequently used or fit for a specific theme or target [18]–[20]. The second type is bottom-up: raw data (e.g., geotagged photos, digital footprints with implicit geographic information or metadata that involved latitude and longitude) to construct a new database or dataset that includes the POI [21]–[25].

By contrast, an AOI might include multiple geographic features or areas with no prominent landmarks, such as a café on a pedestrian street or several neighboring landmarks. Hu et al. [26] proposed that elucidating urban AOIs can provide useful information for city planners, transportation analysts, and supported location-based service providers to plan new businesses and extend existing infrastructure. After they collected Flickr photographic data of six cities in six countries, they used the DBSCAN clustering algorithm to identify urban AOI.

C. Cluster Analysis

Cluster analysis, or unsupervised classification, is one unsupervised learning technique. Cluster analysis can find objects with similar characteristics and can then group homogeneous object into clusters. Each cluster is distinct from the others. This technique is applied widely in fields such as machine learning [27]–[29], image analysis [30]–[31], information retrieval [32]–[33], bioinformatics [34]–[35], and computer graphics.

Major cluster analysis algorithms include the following.

1) Centroid-based Clustering: Centroid-based clustering is an early approach to clustering analysis in which the concept of similarity is computing the distance of a data point from the centroid of the clusters. Based on proximity, objects are assigned to clusters. Typical approaches are K-means and K-medoids. The former, K-means, is the more widely used because it has high-speed performance and easy implementation. However, K-means entails some shortcomings: K-means is sensitive to initial conditions, outliers, etc., and choosing an optimal number is difficult. According to shortcomings of K-means and our dataset, we used X-means for this study to cluster our data, as presented in Section IV.

2) Connectivity-based Clustering (Hierarchical Clustering): Clusters are constructed by calculating "distance" between objects, that can aggregate the similar object into same cluster according to the chosen similarity measure. Similarity measures include the single-linkage agglomerative algorithm, complete-linkage agglomerative algorithm, average-linkage agglomerative algorithm, centroid-linkage agglomerative algorithm, and Ward's minimum variance. In addition, hierarchical clustering is subdivided as explained below.

a) Agglomerative Approach (bottom-up): In this method, each node represents a singleton cluster from the

start. The method proceeds by agglomerating the pair of clusters of minimum dissimilarity to obtain a new cluster. Finally, nodes are merged successively based on their similarities. All nodes belong to the same cluster.

b) Divisive Approach (top-down): All nodes belong to the same cluster. The cluster is classified into sub-clusters, which are divided successively into their own sub-clusters; eventually, each node forms its own cluster. Hierarchical clustering is inappropriate processing for large amounts of data. The final result of this method is presented as a dendrogram. Clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.

3) Density-based Clustering: Density-based clustering can identify distinctive clusters in the data by separating the contiguous region of high point density and regions of low point density. The regions of low point density are typically regarded as noise/outliers. Common examples of density models are DBSCAN [36] and OPTICS [37].

4) Grid-based Clustering: Grid-based clustering quantizes the data space into limited number of cells which form a grid structure, which can obviously reduce the computational complexity, especially for clustering very large datasets. The representative grid-based clustering algorithms are STING [38], WaveCluster [39], and CLIQUE [40].

III. OVERVIEW OF THE METHOD

Figure 1 shows that this section introduces an overview of our method. Our method comprises two components: definition of less-known tourist attractions and data construction.

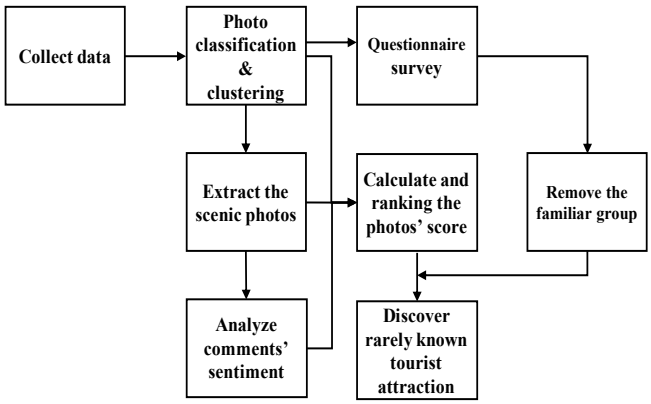


Figure 1. Overview of the method.

A. Definition of Less-Known Tourist Attractions

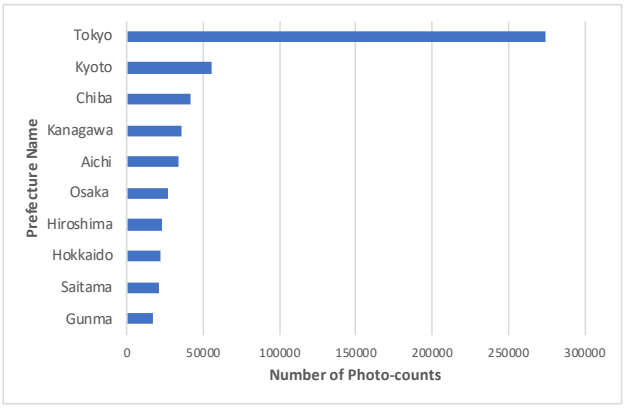
To differentiate well-known and less-known tourist attractions, we adopt two definitions of less-known tourist attractions.

Definition 1: Only some people know about this tourist attraction.

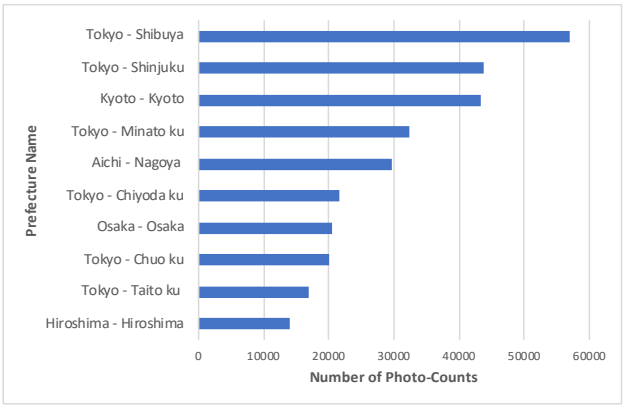
Definition 2: The tourist attraction deserves to be visited. It is attractive for tourists.

B. Data Construction

Using Flickr API, we collected 769,749 photographs taken in 2017 at geolocations throughout Japan. We extracted the photograph latitude and longitude to gather details of addresses through Google geocoding API. We found that 309 photographs were shot in the sky or on the ocean photographs had no details of addresses. We classified these photographs into different prefectures and cities according to the photograph address details. Subsequently, we calculated numbers of photographs of 47 prefectures and 1,158 cities. Figure 2 presents the Top 10 prefectures and cities in terms of the number of photographs.



(a) Top 10 Prefectures for numbers of photographs.



(b) Top 10 Cities for numbers of photographs.

Figure 2. Numbers of photographs.

Next, X-means was used to cluster prefectures and cities into different clusters to administer the questionnaire survey easily. The prefectures are divided into 4 clusters (see Table I). Prefectures are distributed into 14 clusters based on their characteristics. We also defined scores of the prefecture cluster: cluster 1 can yield 4 points, cluster 2 can yield 3 points, and so on. The city cluster score is defined according to questionnaire survey results. Furthermore, we extracted 2,671 scenic photographs with tags that mean scenic in English and Japanese (e.g., "風景", "景色", "scenery"), and collected

these photographs' comments and favorite counts. Then these photographs were ranked using formula proposed in this study. Finally, eliminating familiar city clusters according to result of questionnaire survey that is our final result.

TABLE I. CLUSTERS OF PREFECTURES

Cluster	***Prefectures***
Cluster 1	Tokyo
Cluster 2	Kyoto, Chiba, Kanagawa, Aichi
Cluster 3	Osaka, Hiroshima, Hokkaido, Saitama, Gunma, Nara, Nagano, Okinawa, Hyogo, Fukuoka
Cluster 4	Mie, Tochigi, Shizuoka, Yamanashi, Oita, Okayama, Ibaraki, Aomori, Miyagi, Gifu, Ishikawa, Wakayama, Kagawa, Niigata, Shiga, Ehime, Kumamoto, Akita, Toyama, Fukushima, Nagasaki, Yamagata, Kagoshima, Tottori, Saga, Fukui, Tokushima, Kochi, Yamaguchi, Iwate, Shimane, Miyazaki

IV. SCENIC PHOTOGRAPH EVALUATION

This section presents our approach of photograph evaluation, first illustrating how to detect the characteristic of dataset using a box plot. We can choose the most appropriate method of cluster analysis to classify our dataset. Based on the box plot result, we select X-means to cluster data in this research. We also used the elbow method to set the X-means. After analyzing the positive comments of scenic photographs by application of our formula, weight of our formula is ascertained using the entropy weight method.

A. Box Plot

Before clustering our data, we observe their characteristics. Subsequently, the suitable approach of cluster analysis can be chosen for our data. Therefore, a box plot is used to inspect the four features of Japanese cities: number of photographs of a city, the rate of the number of photographs of a city, the rate of number of photographs of a prefecture, and the average number of photographs of a prefecture.

A box plot, also called box-whisker plot, uses a statistical five number summary of dataset to visualize the data scatter. The five-number summary includes the minimum, first quartile, median, third quartile, and maximum. Moreover, this method is usually used to detect dataset outliers or to assess data symmetry.

After using feature scaling to standardize the Japanese city dataset, the box plot approach was applied to detect this dataset. Results are shown in Figure 3, which presents all cities' data scattering. A disparity between our dataset and the outliers are readily apparent in these data. These outliers cannot be eliminated because each datum represents a Japanese city. Less-known tourist attractions might exist in this city. Therefore, improper clustering methods that are sensitive to outliers should be avoided. For this study, we used X-means to cluster our data as we introduce with the next subtask.

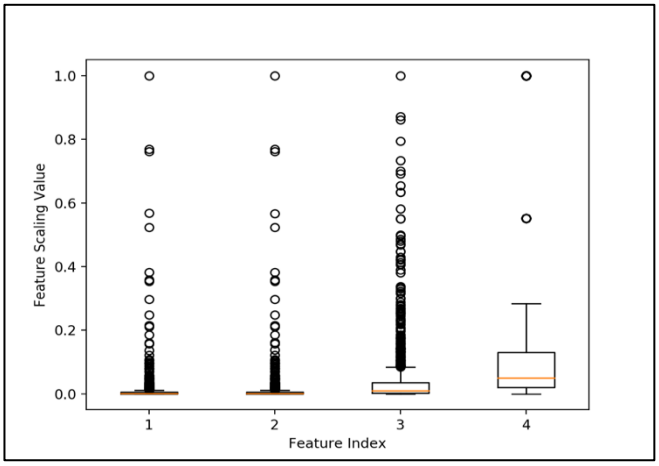


Figure 3. Box plot of dataset. Index 1 is the number of photographs of the city. Index 2 is the rate of the number of photographs of city. Index 3 is the rate of the number of photographs of the prefecture. Index 4 is the average of the number of photographs of the prefecture.

B. X-means Algorithm

The X-means algorithm is one clustering technique proposed by Pelleg and Moore [41] to improve the shortcomings of K-means. According to the BIC score, the X-means algorithm can automatically determine the optimum number of clusters that user set only minimum and maximum of clusters. Additionally, this approach greatly reduces the probability of being trapped into a local optimum and using the kd-tree to increase the computational speed.

The process steps of X-means are presented below.

1. Input dataset, setting the minimum (K_{min}) and maximum (K_{max}) parameters for the number of clusters (K).
2. Run K-means. (K= K_{min})
3. Run 2-means in each cluster according to the BIC score to decide splitting it or not.
4. If $K > K_{max}$, then stop and report the best scoring model found during the search. Otherwise, go to step 2.

Considering the outliers existing in the data, we used this method to distribute the 47 prefectures and 1,158 cities into different clusters according to their respective characteristics. Furthermore, we set the minimum cluster of X-means by referring to the result of elbow method. Finally, we obtain more scattered results than those obtained earlier by X-means. The city cluster result is presented in Table II.

TABLE II. RESULT OF CITY CLUSTER

Cluster	City counts	Percentage
Cluster 1	89	8%
Cluster 2	24	24%
Cluster 3	254	22%
Cluster 4	3	0%
Cluster 5	5	0%
Cluster 6	88	8%
Cluster 7	84	7%
Cluster 8	4	0%
Cluster 9	52	4%
Cluster 10	42	4%
Cluster 11	20	2%
Cluster 12	39	2%
Cluster 13	126	11%
Cluster 14	328	28%
Total	1,158	100 %

C. Elbow Method

The elbow method is the most popular technique used to ascertain the optimum number of clusters (*K*). The elbow method concept is calculating the total within-cluster sum of squares (wss) for each number of clusters and plotting the curve of wss. Therefore, we can ascertain the optimum number of clusters by finding the location of the warp (elbow point) in the plot of the elbow method.

Figure 4 and Figure 5 present results of the elbow method for cities and prefectures. The values of *K* are shown on the X axis. Those of wss of each *K* are shown on the Y axis. Moreover, in Figure 5, although the wss falls rapidly with *K* increasing from 1 to 4, the slope of line still has a marked change thereafter. After *K*=9, the curve goes down very slowly. Consequently, we determine optimum number of clusters as 9.

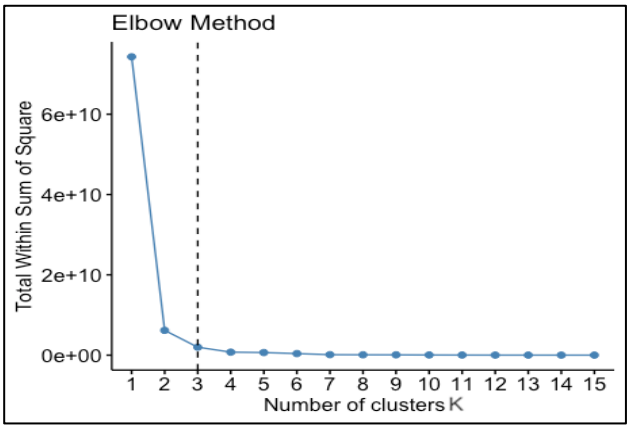


Figure 4. Elbow method of prefecture data.

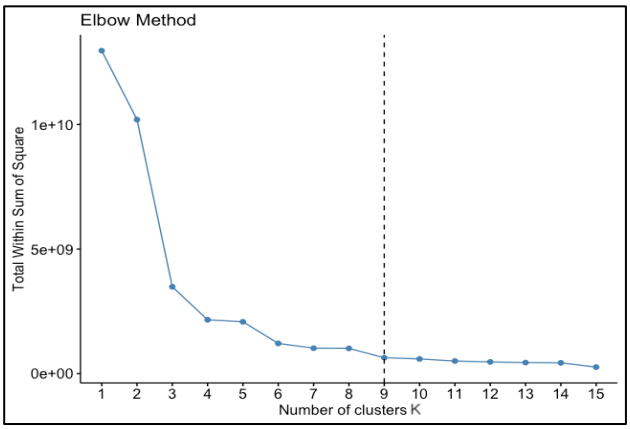


Figure 5. Elbow method of city data.

D. Comments' Sentiment

For this subtask, we assume that the positive comment is that of factors to ascertain whether this sightseeing spot is attractive for tourists to visit or not. Therefore, we collected the scenic photograph comments. Additionally, we eliminated the owner comments from the total comments because almost all of these comments are merely replies to the viewer comments. Subsequently, we analyzed these comments and extracted the positive comments, as shown in Table III.

TABLE III. NUMBER OF POSITIVE COMMENTS

	Viewer comments	Owner comments	Sum
Positive comments	1,602	248	1,850
Total comments	2,417	572	2,989

We specifically examined English and Chinese comments using TextBlob [42] and SnowNLP [43], which yielded the score of sentiment representing the probability of positive meaning. Scores of English comments' sentiments were -1 to 1. The Chinese sentiment scores were 0 to 1. To increase the accuracy of judgment, the score of English positive comments was assumed as more than 0.3; the scores of Chinese positive comments were assumed to be greater than 0.4.

E. Formula of Evaluation

Considering the definitions of less-known tourist attractions and data construction, we propose a formula to calculate the score S_i to rank the photographs.

$$S_i = \sum_{p=1}^{3} F_{pi} W_p + R_i \,, 0 < W_p < 1 \; and \; \sum_{p=1}^{3} W_p = 1 \qquad (1)$$

In equation (1), F_{1i} represents the prefecture cluster point; W_1 is F_{1i} 's weight. F_{2i} represents a city cluster point; W_2 is the weight associated with F_{2i} . F_{3i} represents the photographs' favorite counts. W_3 is F_{3i} 's weight. R_i represents the positive comment count of the photographs,

TABLE IV. PART OF JAPANESE RANKING RESULT

Address	Neighboring tourist attraction	Prefecture score	City score	Favorites	Positive comments	Score
2871 Onna, Onna-son Kunigami-gun, Okinawa, 904-0411, Japan	Resort	2	2.22	1548	56	1108.73
Yunohama hotel, 1-2-30, Yunokawacho, Hakodate-shi, Hokkaido, 042-0932, Japan	Hot spring street	2	2.02	337	13	241.83
14-16 Suehirocho, Hakodate-shi, Hokkaido, 040-0053, Japan	Kanemori Red Brick Warehouse	2	2.02	306	5	219.51
510 Tangocho Takano, Kyotango-shi, Kyoto, 627-0221, Japan	---	3	1.59	187	4	134.49
Kendou 388sen, Inuma, Kawanehon-cho Haibara-gun, Shizuoka, 428-0402, Japan	---	1	1.4	126	46	91.27
Sinkawagensi 58, Fukuoka Yatsumiya, Shiroishi-shi, Miyagi, 989-0733, Japan	---	1	1.64	123	2	88.37

TABLE V. PART OF TAIWANESE RANKING RESULT

Address	Neighboring tourist attraction	Prefecture score	City score	Favorites	Positive comments	Score
Ryuuanzi, Ryoanji Goryonoshitacho, Ukyo-ku Kyoto-shi, Kyoto, 616-8001, Japan	Temple of the Dragon at Peace	3	3.74	100	6	74.17
156 Fumoto, Fujinomiya-shi, Shizuoka, 418-0109, Japan	---	1	1.63	99	32	73.35
1070 Kodachi, Minamitsuru Gun Fujikawaguchiko Mac, Yamanashi, 401-0302, Japan	Lake Kawaguchi	1	1.89	94	19	69.5
Motosumichi, Minamigeuma-gun, Yamanashi Prefecture, Japan	---	1	1.4	92	27	68.11
Kendou60sen, Tazawako Tazawa, Semboku Shi, Akita, 014-1204, Japan	Lake Tazawa	1	1.69	69	36	51.48
86 Himata, Toyama Shi, Toyama, 930-0912, Japan	---	1	1.69	67	5	49.47

which is processed by feature scaling. In this formula, R_i is regarded as an additional score because most photographs have no associated comments. The weight of R_i is almost equal 0. The photograph favorite counts and positive comments were assumed as factors attracting someone to visit. Therefore, all scenic photographs can be ranked using this formula, as shown in Table IV and Table V.

Table IV and Table V present some Taiwanese and Japanese ranking results. The first column is the GPS address of the photograph from Google API. The second column is the neighboring popular tourist attraction. The third and fourth columns are photograph cluster scores. The fifth column shows the favorite count of photographs. The sixth column shows counts of the photograph positive comments. The last column presents the photograph score as calculated using our formula. A high score indicates that the place is attractive to travelers. In Table IV, the address of the first row is a famous resort in Okinawa. The second row presents a hotel on a famous hot spring street. The third row is a well-known tourist attraction in Hokkaido. In Table V, the first row presents a renowned and historical temple in Kyoto. The third row location is near Lake Kawaguchi: one of the Fuji Five Lakes. The place of fifth row is near Lake Tazawa, the deepest lake in Japan. Others are obscure places.

F. Entropy Weight Method (EWM)

For this study, we used EWM to set the weights used for the formula. EWM is an objective set weight method because it depends only on the discreteness of data. Actually, EWM is used widely in the fields of engineering, socioeconomic studies, etc. [44]–[46].

In information theory, entropy is a kind of uncertainty measure. When information is greater, uncertainty and entropy are smaller. Based on the entropy information properties, one can estimate the randomness of an event and the degree of randomness through calculation of the entropy value. Furthermore, entropy values are used to gauge a sort of discreteness degree of index. When the degree of discreteness is larger, the index affecting the integrated assessment is expected to be greater.

To complete the setting of the formula weights, we require the steps, as described below.

1) Calculate the ratio (P_{ij}) of the i-th index under the j-th index. Therein, x_{ij} denotes the j-th index of the i-th sample.

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, (i = 1, \dots, n;\ j = 1, \dots, m) \quad (2)$$

2) *Calculate the entropy value (e_j) of the j-th index.*

$$e_j = -k \sum_{i=1}^{n} P_{ij} \ln\left(P_{ij}\right), (j = 1, \dots, m;\ k = \frac{1}{\ln(n)} > 0) \quad (3)$$

3) *Calculate the discrepancy of information entropy (d_j).*

$$d_j = 1 - e_j, (j = 1, \dots, m) \quad (4)$$

4) *Calculate the weight (w_i) of each index.*

$$w_i = \frac{d_j}{\sum_{j=1}^{m} d_j}, (j = 1, \dots, m) \quad (5)$$

We analyzed the prefecture cluster score (F_{1i}), the city cluster score (F_{2i}), and the favorite counts (F_{3i}) of 2,671 scenic photographs. Results show the weight of the formula in this research by EWM. In the equations (1), Taiwanese and Japanese weights differ because their city clusters are assigned distinct scores based on the results of questionnaire surveys. The weight results are shown in Table VI: Taiwanese W_1 is equal to 0.1338; W_2 is equal to 0.1346 and W_3 is equal to 0.7316. Japanese W_1 is equal to 0.1619; W_2 is equal to 0.1228 and W_3 is equal to 0.7152.

TABLE VI. TAIWANESE AND JAPANESE WEIGHTS

	W_1	W_2	W_3
Taiwanese weight	0.1338	0.1346	0.7316
Japanese weight	0.1619	0.1228	0.7152

V. LESS-KNOWN TOURIST ATTRACTION ESTIMATION

A. Familiarity Level of Japanese City

For this study, we assume the less-known tourist attractions might be included in unfamiliar city clusters. Accordingly, a questionnaire was designed and administered to 115 Taiwanese and 123 Japanese people to ascertain their level of familiarity with Japanese cities. Nevertheless, surveying levels of familiarity of each city (1,158 cities) from respondents was difficult. For that reason, we clustered the Japanese city data. Thereby, we were able to select a city's name randomly from each cluster to decrease the number of questionnaire questions. It was easier to find which cities were unfamiliar to respondents.

According to the scale of each cluster, 30 city names were selected randomly for this questionnaire. Participants were provided with five choices to answer the city questions: (1) I have absolutely no idea. (2) I have heard of this city, but I do not know its tourist attractions. (3) I have heard of this city and know its tourist attractions. (4) I have been to this city, but I do not know its tourist attractions. (5) I have been to this city and know its tourist attractions. A respondent choosing option (1) is assigned 1 point for this question; option (2) yields 2 points, and so on, with higher scores representing greater familiarity with this city.

Considering that we used the survey sampling approach to conduct the questionnaire survey, it might include sampling error. To decrease inaccuracy from the sampling error, we categorized the cluster as a less-known one using *t*-tests and *p*-value.

$$\mathrm{t} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (6)$$

Student's t-test can determine whether a statistically significant difference exists between the means of two unrelated groups. This approach has three types: one-sample t-test, independent-sample t-test, and paired sample t-test. For this study, we used one-sample t-test to analyze our result of questionnaire survey, which can compare population means with a sample mean, and found their relation. In the following equation (6), $\overline{x}$ represents the sample means, μ denotes the population mean, s stands for the sample standard deviation, and n is the sample size. After calculating the t-test values, we used *p*-value to determine whether the sample mean was greater than the population mean or not. If the *p*-value of cluster was less than 0.05, then we judged this cluster as an unfamiliar cluster. Conversely, the cluster will be categorized into familiar clusters when the *p*-value is greater than 0.05.

TABLE VII. TAIWANESE UNFAMILIAR CITY CLUSTERS

Cluster	Sample mean	*t*-test value	*p*-value	Unfamiliar
Cluster 1	1.63	-1.40	0.08	
Cluster 2	1.89	2.05	0.98	
Cluster 3	1.40	-7.63	0.00	✓
Cluster 4	3.74	18.07	1.00	
Cluster 5	2.90	8.55	1.00	
Cluster 6	1.70	0.12	0.55	
Cluster 7	1.56	-2.50	0.01	✓
Cluster 8	2.55	6.49	1.00	
Cluster 9	1.74	0.62	0.73	
Cluster 10	1.69	-0.06	0.47	
Cluster 11	1.59	-1.11	0.13	
Cluster 12	1.52	-2.82	0.00	✓
Cluster 13	1.37	-7.22	0.00	✓
Cluster 14	1.31	-11.33	0.00	✓
Population mean	1.69	---	---	---

TABLE VIII. JAPANESE UNFAMILIAR CITY CLUSTERS

Cluster	Sample mean	*t*-test value	*p*-value	Unfamiliar
Cluster 1	2.02	0.22	0.59	
Cluster 2	2.96	7.14	1.00	
Cluster 3	1.61	-7.24	0.00	✓
Cluster 4	4.47	29.79	1.00	
Cluster 5	3.51	10.23	1.00	
Cluster 6	1.80	-3.53	0.00	✓
Cluster 7	1.47	-11.01	0.00	✓
Cluster 8	3.24	9.55	1.00	
Cluster 9	2.32	3.46	1.00	
Cluster 10	2.22	2.28	0.99	
Cluster 11	1.59	-4.42	0.00	✓
Cluster 12	2.09	0.88	0.81	
Cluster 13	1.64	-6.50	0.00	✓
Cluster 14	1.40	-15.37	0.00	✓
Population mean	2.01	---	---	---

Table VII and Table VIII show that we calculated the average scores of respective clusters. Table VII and Table VIII present results of unfamiliar clusters. The first column shows the number of clusters. The second column is each cluster average score from the questionnaire survey. The third column shows the t-test statistic value. The fourth column presents p-values. The last column presents which cluster is unfamiliar. In Table VII, one can understand that cluster 3, cluster 7, cluster 12, cluster 13, and cluster 14 are unfamiliar to Taiwanese. Moreover, Table VIII shows that Japanese people are unfamiliar with cluster 3, cluster 6, cluster 7, cluster 11, cluster 13, and cluster 14. Finally, we can remove these familiar clusters from the ranking results of Section IV as our aim.

B. *Verification Experiment*

Based on the discussion presented above, we can ascertain which group is unfamiliar to the Taiwanese respondents (clusters 3, 7, 12–14) and to the Japanese respondents (clusters 3, 6, 7, 11, 13, 14). In this section, we also use the questionnaire to verify these less-known tourist attractions, which are obscure but attractive to respondents.

For the verification experiment, we extracted the top 10 less-known tourist attractions from nine cities of the Taiwanese and Japanese unfamiliar clusters to investigate 10 Taiwanese people (who have touristic experience in Japan) and 10 Japanese people, whose questionnaires responses were dissimilar. Two questions were asked for each attraction: "Do you know this city?" If respondents probably know this city, then the answer was "Yes." The second question was "According to this photograph, do you want to visit this place of city?" For the second question, respondents assigned a score of 1–5 for the attraction, with a higher score indicating greater attraction.

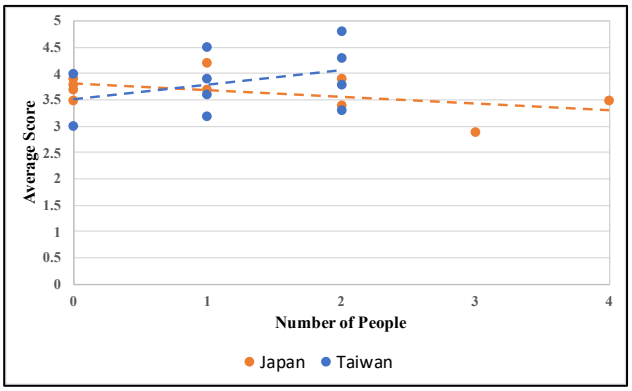


Figure 6. Result of verification experiment.

This result of the verification experiment showed that these places are known by extremely few people, which is better than the previous result. In Figure 6, each point represents a place in the questionnaire, the horizontal axis presents how many people know the less-known tourist attraction, the vertical axis shows the attractive level of each less-known tourist attraction. The Taiwanese result presents the average score of four places are over than 4 points, which means these places are attractive for Taiwanese respondents. Especially, one place score approaches the full mark. Some Taiwanese respondents reported that a few scenic photographs are similar to scenery in their own country, which might influence their decision. Moreover, for the place with the lowest score, the photograph quality was not very high. As a result, the respondents assigned few points for this place. The required cost of Taiwanese includes a monetary cost and time cost, which are higher than those of Japanese people. Consequently, Taiwanese prefer to choose tourism attractions that include local characteristics or exceptional landscapes.

For the Japanese result, although only one spot yielded over 4 points, two other spots yielded over 3.5 points, indicating that Japanese respondents are not excluded from visiting these spots. Furthermore, we investigated the answers of each Japanese respondent in depth and detected that the disparity between their decisions decreased the average. This situation expresses that respondents chose the answer according to their preference of scenic spots. For instance, someone who likes the ocean, but does not like mountains will assign a higher score for seascape photographs. Therefore, the average of each spot is less than 4 points. Figure 6 also shows an interesting situation. The Taiwanese trendline shows that the number of people and average score are in direct proportion, but the Japanese trendline is inverse to that of the Taiwanese curve: Japanese people prefer to visit less-known tourist attractions.

VI. DISCUSSION AND CONCLUSION

We proposed a novel method to identify less-known tourist attractions for people of different nationalities. By collecting and analyzing Flickr photograph information, we classified them into prefectures and cities. Subsequently, we classified these prefectures and cities into different groups.

Additionally, we used a questionnaire to survey Taiwanese respondents and Japanese respondents. We obtained unfamiliar city clusters of Taiwanese and Japanese respondents. Scenic photographs were ranked using the formula for this research. Familiar city clusters were removed from respondent ranking results. A second questionnaire survey verified our results. Through this research, we found less-known tourist attractions for travelers.

The first questionnaire survey gave the surprising result that Taiwanese respondents are more familiar with Japanese cities than Japanese respondents are. The reason might be that Taiwan and Japan are neighboring countries. In addition, air travel from Taiwan to Japan is cheaper, which might engender a higher frequency of Taiwanese taking trips to Japan. Results show that most Taiwanese respondents prefer individual travel in Japan to travelling with groups.

The verification experiment revealed an interesting thing: we provide two seascape photographs from distinct spots for Taiwanese respondents. One photograph shows the "torii", which is the traditional gate of Japanese shrines. The other only has a clear ocean and beach. Two Taiwanese respondents said that the seascape is common in Taiwan, but they are very interested in the first seascape because of this spot, which includes "torii." Scenic photographs including some special landmarks are expected to increase the attractiveness of these spots.

Interviews of some Taiwanese respondents to ascertain what factors lead them to prefer to travel in Japan indicated four main reasons it is attractive to Taiwanese. The first reason is that air tickets are cheaper and the flight time is short. The second reason is that the Japanese environment is neat and tidy. Furthermore, public security is high. The third reason is that Japanese foods are delicious and exquisite. The fourth reason is that Japanese character and culture are similar to those of Taiwan, which can help Taiwanese people travel in Japan easily.

As future work, after collecting and analyze more photographs taken in distinct years, we expect to sort the photographs with lowest quality from our data and remove them. Providing higher-quality photographs for travelers might induce them to visit. Considering more factors of discovering less-known tourist attractions, we expect to improve the formula used for this research. Less-known tourist attractions will be classified into different types (e.g., ocean, mountain, sky), seasons, weather, days, and nights according to the times and contents of photographs. We also want to provide a personal recommendation service based on collaborative filtering.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 19K20418.

REFERENCES

[1] J. Lin, S. Wen, M. Hirota, T. Araki, and H. Ishikawa, "Analysis of Rarely Known Tourist Attractions by Geo-tagged Photographs," MMEDIA, Mar. 2019.

[2] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera, "Personalized Tour Recommendation Based on User Interests and Points of Interest Visit Durations," IJCAI, pp. 1778–1784, Jul. 2015.

[3] I. Memon, L. Chen, A. Majid, M. Lv, I. Hussain, and G. Chen, "Travel Recommendation Using Geo-tagged Photos in Social Media for Tourist," International Journal of Wireless Personal Communications, vol. 80, pp. 1347–1362, Feb. 2015.

[4] S. Jiang, Z. Qian, T. Mei, and Y. Fu, "Personalized Travel Sequence Recommendation on Multi-Source Big Social Media," IEEE Trans. Big Data, vol. 2, no. 1, pp. 43–56, Mar. 2016.

[5] X. Peng and Z. Huang, "A Novel Popular Tourist Attraction Discovering Approach Based on Geo-Tagged Social Media Big Data," ISPRS International Journal of Geo-Information, vol. 6, no. 7, pp. 216, Jul. 2017.

[6] A. Hausmann et al. "Social Media Data Can be Used to understand Tourists' Preferences for Nature-Based Experiences in Protected Areas," Conservation Letters, vol. 11, no. 1, Jan. 2017.

[7] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," ACM Multimedia, pp. 589–598, Nov. 2012, ISBN: 978-1-4503-1089-5.

[8] C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa, "Discovering Obscure Sightseeing Spots by Analysis of Geo-tagged Social Images," ASONAM, pp. 590–595, Aug. 2015, ISBN: 978-1-4503-3854-7.

[9] S. F. Schubert, J. G. Brida, and W. A. Risso, "The impacts of international tourism demand on economic growth of small economies dependent on tourism," Tourism Management, vol. 32, pp. 377–385, Apr. 2011.

[10] A. Konstantinos, "Scale of hospitality firms and local economic development-dividence from Crete," Tourism Management, vol. 23, pp. 333–341, Aug. 2002, ISSN 0261-5177.

[11] R. R. Croes, "A paradigm shift to a new strategy for small island economies: embracing demand side economics for value enhancement and long term economic stability," Tourism Management, vol. 27, pp. 453–465, Jun. 2006.

[12] F. Michael, "Tourism as a feasible option for sustainable development in small island developing states (SIDS): Nauru as a case study," Pacific Tourism Review, vol. 3, no. 2, pp. 133–142(10) ,1999.

[13] B. Lin and H. Liu, "A study of economies of scale and economies of scope in Taiwan international tourist hotels," Asia Pacific Journal of Tourism Research, vol. 5, pp. 21–28, Apr. 2007.

[14] G.I. Crouch and J.R.B. Ritchie,"Tourism, Competitiveness, and Societal Prosperity," vol. 44, pp. 137–152, Mar. 1999.

[15] B. Algieri, A. Aquino, and M. Succurro, "International competitive advantages in tourism: An eclectic view," Tourism Management Perspectives, vol. 25, pp. 41–52, Jan. 2018.

[16] K. Kakamu, W. Polasek, and H. Wago, "Spatial interaction of crime incidents in Japan," Mathematics and Computers in Simulation, vol. 78, no. 2, pp. 276–282, Jul. 2008.

[17] D. Altindag, "Crime and International Tourism," Journal of Labor Research, vol. 35, no. 1, pp. 1–14, Mar. 2014, doi: 10.1007/s12122-014-9174-8.

[18] H. Chuang, C. Chang, T. Kao, C. Cheng, Y. Huang, and K. Cheong, "Enabling maps/location searches on mobile devices: Constructing a POI database via focused crawling and information extraction," Int. J. Geogr. Inf. Sci., vol. 30, pp. 1405–1425, Jan. 2016.

[19] D. Jonietz and A. Zipf, "A. Defining fitness-for-use for crowdsourced points of interest (POI)," ISPRS International Journal of Geo-Information, vol. 5, Aug. 2016.

[20] A. Rousell, S. Hahmann, M. Bakillah, and A. Mobasheri. "Extraction of landmarks from OpenStreetMap for use in

navigational instructions," In Proceedings of the 18th AGILE International Conference on Geographic Information Science, Jun. 2015.

[21] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring Millions of Footprints in Location Sharing Services," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 81–88, Jun. 2011.

[22] E. Spyrou, M. Korakakis, V. Charalampidis, A. Psallas, and P Mylonas. "A Geo-Clustering Approach for the Detection of Areas-of-Interest and Their Underlying Semantics," Algorithms, Mar. 2017, DOI:10.3390/a10010035.

[23] A. Skovsgaard, D. Idlauskas, and C. S. Jensen. "A clustering approach to the discovery of points of interest from geo-tagged microblog posts," 2014 IEEE 15th International Conference on Mobile Data Management (MDM), pp. 178–188, Jul. 2014, DOI: 10.1109/MDM.2014.28.

[24] D. D Vu, H. To, W. Shin, and C. Shahabi. "GeoSocialBound: An Efficient Framework for Estimating Social POI Boundaries Using Spatio–Textual Information," Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, Jun. 2016, DOI:10.1145/2948649.2948652.

[25] C. Kuo, T. Chan, I. Fan, and A. Zipf, "Efficient Method for POI/ROI Discovery Using Flickr Geotagged Photos," ISPRS International Journal of Geo-Information, Mar. 2018, DOI:10.3390/ijgi7030121.

[26] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li, and S. Prasadd, "Extracting and understanding urban areas of interest using geotagged photos," Computers, Environment and Urban Systems, vol. 54, pp. 240–254, Nov. 2015.

[27] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," International Conference on Machine Learning (ICML), pp. 478–487, 2016.

[28] L. Wang, "Discovering phase transitions with unsupervised learning," Phys. Rev. B 94, Nov. 2016, DOI: 10.1103/PhysRevB.94.195105.

[29] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for Internet of Things data analysis: A survey," Digital Communications and Networks, vol. 4, pp. 161–175, Aug. 2018.

[30] N. Dhanachandra, Y. J. Chanu, and K. M. Singh, "Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm," Procedia Computer Science, vol. 5, pp. 764–771, 2015.

[31] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Performance analysis of image segmentation using watershed algorithm, fuzzy C-means of clustering algorithm and Simulink design," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Mar. 2016.

[32] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," arXiv preprint arXiv:1707.02919, 2017.

[33] W. B. A. Karaa, A. S. Ashour, D. B. Sassi, P. Roy, N. Kausar, and N. Dey, "MEDLINE Text Mining: An Enhancement Genetic Algorithm Based Approach for Document Clustering," Applications of Intelligent Optimization in Biology and Medicine, pp. 267–287, Mar. 2015.

[34] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective," Trends in Food Science & Technology, vol. 72, pp. 83–90, Feb. 2018.

[35] M. D. M. Fernández-Arjona, J. M. Grondona, P. Granados-Durán, P. Fernández-Llebrez, and M. D. López-Ávalos "Microglia Morphological Categorization in a Rat Model of Neuroinflammation by Hierarchical Cluster and Principal Components Analysis," Front Cell Neurosci., vol. 11, Aug. 2017, DOI: 10.3389/fncel.2017.00235.

[36] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231, 1996.

[37] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, vol. 28, pp. 49–60, 1999.

[38] W. Wang, J. Yang, and R. R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," In Proceedings of the 23rd International Conference on Very Large Data Bases, pp. 186–195, Aug. 1997, ISBN:1-55860-470-7.

[39] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases," In Proceedings of the 24rd International Conference on Very Large Data Bases, pp. 428–439, Aug. 1998, ISBN:1-55860-566-5.

[40] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pp. 94–105, Jun. 1998, ISBN:0-89791-995-5.

[41] D. Pelleg and A. W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," In Proceedings of the 17th International Conf. on Machine Learning, pp.727–734, Jul. 2000.

[42] TextBlob.[Online]. Available from: https://pypi.org/project/textblob/ 2019.08.15

[43] SnowNLP.[Online].Available from: https://pypi.org/project/snownlp/ 2019.08.15

[44] Y. He, H. Guo, M. Jin, and P. Ren, "A linguistic entropy weight method and its application in linguistic multi-attribute group decision making," Nonlinear Dynamics, vol. 84, no. 1, pp. 399–404, Jan. 2016.

[45] Y. Ji, G. H. Huang, and W. Sun, "Risk assessment of hydropower stations through an integrated fuzzy entropy-weight multiple criteria decision making method: A case study of the Xiangxi River," Expert Systems with Applications, vol. 42, no. 12, pp. 5380–5389, Jul. 2015.

[46] A. Delgado and I. Romero, "Environmental conflict analysis using an integrated grey clustering and entropy-weight method: A case study of a mining project in Peru," Environmental Modelling & Software, vol. 77, pp. 108–121, Mar. 2016.

EPOS: European Plate Observing System: Challenges being addressed

Keith G. Jeffery
Keith G. Jeffery Consultants
Faringdon, UK
Email: keith.jeffery@keithgjefferyconsultants.co.uk

Daniele Bailo
EPOS-ERIC Office
Istituto Nazionale di Geofisica e Vulcanologia
Rome, Italy
Email: daniele.bailo@ingv.it

Kuvvet Atakan
Department of Earth Science
University of Bergen
Bergen, Norway
Email: kuvvet.atakan@uib.no

Matt Harrison
Director Informatics
British Geological Survey
Keyworth, UK
Email: mharr@bgs.ac.uk

***Abstract*—The European plate observing system (EPOS) addresses the problem of homogeneous access to heterogeneous digital assets in geoscience of the European tectonic plate. Such access opens new research opportunities. Previous attempts have been limited in scope and required much human intervention. EPOS adopts an advanced Information and Communication Technologies (ICT) architecture driven by a catalog of rich metadata. The novel architecture together with challenges encountered and solutions adopted are presented.**

Keywords-geoscience; information; metadata; CERIF; distributed databases; research infrastructures

I. INTRODUCTION

This paper is an extended and improved version of that presented at the GeoProcessing 2019 conference [1] and details the current challenges being addressed.
First, we introduce the challenges that have faced the EPOS project and cover briefly previous relevant work.

A. Overview

Information pertaining to geoscience in Europe is heterogeneous in language, structure, semantics, granularity, content precision and accuracy, method of collection and more. However, there is an increasing demand for access to and utilisation of this information for decision-making in industry and government policy. EPOS is providing a mechanism for homogeneous access to - and comfortable utilisation of - this base of rich heterogeneous assets.

EPOS may be considered a journey. During the EPOS Preparatory Project (EPOS-PP) domain communities discovered their commonality and differences and – particularly - their digital assets offered as Thematic Core Services (TCSs). This process was lengthy, requiring much interaction to understand similarities and differences including in the use of language to describe requirements and offered assets. The whole process was facilitated by the EPOS ICT team. The assets were documented in a database, which demonstrated clearly (a) that considerable assets existed (more than 400); (b) that the organizations (covering more than 250 research infrastructures (RIs)) owning the digital assetswere willing to make them available (sometimes subject to conditions); (c) that there was overlap of assets between some communities; (d) that multidisciplinary geoscience could be achieved by providing appropriate interoperation mechanisms to make the assets available to all. An extensive review of possible architectural solutions across many sectors of research, government and industry was conducted but none satisfied the requirements. A novel, leading-edge architecture was proposed, discussed and agreed among the TCSs and the ICT team. This was then implemented as a prototype to demonstrate that, indeed, interoperation across heterogeneous communities and their assets could be achieved.

The task of the EPOS Implementation Project (EPOS-IP) is to build a geoscience environment (including governance, legal, financial, training and social aspects as well as technical ICT contributions) for the community. This Version 1.0 of the EPOS platform will then be maintained and extended by the EPOS European Research Infrastructure Consortium (EPOS-ERIC), the legal body set up by the supporting Member States providing greater sustainability for maintenance, coordination and access into the future.

There are currently 10 different TCS communities (with an additional two pending approval) with distinct and variable but complementary coverage over the entire

spectrum of solid Earth sciences. While some of the TCSs are discipline specific such as seismology, geodesy, geomagnetism, geology, others are more cross-disciplinary in their origin such as near-fault observatories, volcano observations, satellite observations of geohazards, anthropogenic hazards, multi-scale laboratories and geo-energy test-beds for low-carbon energy. Many of the assets are based on measurements by sensors or laboratory equipment covering many aspects of physics and chemistry. TCSs have variable histories of developments where some have longer history (>100 years) and hence are more mature than the others. They have established their own distinctways of working, data and software specifications. They have local domain-specific standards (although some are International or European) and constraints especially relating to their interoperation with other International organisations in their specific domain. A critical issue is the harmonisation of the descriptions of the TCSs' assets from their own local metadata standards (currently 17 different standards) as a single rich canonical metadata format with formal syntax (structure) and declared semantics (meaning of terms used). The intention is to assist interoperation of the TCSs assets within and between communities by means of the Integrated Core Services (ICS) – including the rich metadata catalog – which forms the entry-point to EPOS and the view over the EPOS assets made available withinthe TCSs.

The key requirements are as follows:

1. Minimal interference with existing communities' operations and developments including IT;
2. Easy-to-use user interface;
3. Access to assets through a metadata catalog: initially services but progressively also datasets, workflows, software modules; computational facilities, instruments/sensors all with associated organisational information including persons in roles such as experts and service managers;
4. Progressive assistance in composing workflows of services, software and data to deploy on e-Infrastructures to achieve research infrastructure user objectives.

B. Interoperability Challenge

EPOS comprises 10 communities of users characterised by domain of interest (TCSs), which supply the metadata describing the assets to the ICS. These communities have varying levels of expertise in the use of ICT for their scientific domain. The processing techniques used vary from domain to domain. With differing domains, the data models used for data collection and processing, and the metadata associated with associated services, equipment and that data, vary greatly. Across many domains geo-coordinates (including both space and time) are common, but not necessarily using the same coordinate system not standard for representation. Similarly, there are multiple metadata standards used for descriptive keywords and other attributes.

The software used for processing in each community is different, although there is some commonality, e.g., where several communities use satellite imagery. The data processing methods – from validating raw data, summarising, analytics, simulation and visualisation – varies from community to community. The more advanced communities have sophisticated workflows integrating data and processing with advanced computing facilities addressing key scientific challenges with big-data analyses and modelling. However, this is a fast-changing field and while workflows used systems like Taverna [2] in the past, the current favourite is Jupyter Notebooks [3], [4], [5]. Similarly, previous use of high-performance computers under the PRACE [6] umbrella is changing to use of commercial Cloud Computing services (such as Amazon) or EOSC (European Open Science Cloud) [7].

Most of the domains have organised computing and observational (sensor-networks) infrastructure for their purposes at institutional, national and trans-European levels. However, additionally it may be necessary to utilise supercomputing facilities, which require procurement or agreements for use as well as mechanisms to deploy the processing workflow. Progressively, EPOS is working more closely with European Open Science Cloud (EOSC) to provide such facilities, although the EPOS architecture is designed to be independent of e-Infrastructure.

e-Is (e-Infrastructures) continue to provide a level of services common to – and used by – many Research Infrastructures (RIs) and other research environments. The major e-Infrastructures of relevance to EPOS-IP are:

1. GEANT: the academic network in Europe, which brings together the national computational networks [8];
2. EGI: a foundation and organisation providing infrastructure computing and data facilities for research [9];
3. EUDAT an EC-funded project to provide infrastructure services for datasets including curation, discovery [10];
4. PRACE: a network providing resources on supercomputers throughout Europe [6];
5. EOSC: the European Open Science Cloud, which aims to provide infrastructure services for research with the first pilot project starting in January 2017 [7] and subsequently the EOSC-Hub, which is soliciting services;
6. OpenAIRE: an EC-funded project to provide metadata to access research publications and – started recently – related datasets [11].

Participant organisations in EPOS have been involved to a greater or lesser extent in all of these activities. In particular EPOS TCSs (with support from the ICS team) have been conducting pilot projects with EGI, PRACE and EUDAT and EPOS is involved in the EOSC pilot.

The level of expertise in both the science and the use of IT varies from community to community. There has been quite some education effort from the central IT team towards the domain communities to explain current computing techniques – especially for cross-domain interoperability, which previously had not been a consideration.

C. Previous Work

EPOS provides an original approach to the provision of homogeneous access over heterogeneous digital assets. Previous work has been within a limited domain (where standards for assets and their metadata may be consensual thus reducing heterogeneity) and involving much manual intervention with associated costs and potential errors. An early attempt for geoscience information was Filematch [12], which exhibited those problems. NASA has a Common Metadata Repository (CMM). In 2013 NASA decided it could not persuade every data provider to use ISO19115 so developed the Unified Metadata Model (UMM) [13] to and from which other metadata standards are converted. This follows the approach used in EPOS already and provides some assurance of the direction being taken. The Open Geoscience Consortium (OGC) has produced a series of standards. GeoNetwork [14] has established a suite of software based around the OGC ISO19115 metadata standard; however, despite its open nature this software 'locks in' the developer to a particular way of processing and does not assist in the composition and deployment of workflows and the metadata is insufficiently rich for automated processing. Some major projects run parallel to EPOS: EarthCube [15] is a collection of projects providing designs and tools for geoscience including interoperability in USA, which investigated the brokering approach – encountering the 'explosion problem' of many bilateral brokers and is now following a metadata-driven brokering mechanism like that used in EPOS, which reduces the number of convertors for metadata from (n(n-1)) to n; Auscope [16] is a set of related programmes in Australia with one (AuScope GRID) providing access to assets and using ISO19115 as the metadata standard with the deficiencies mentioned above; GEOSS [17] is developing interoperation through a system or systems approach, which naturally requires many bilateral interfaces to be maintained with consequent difficulties and maintenance costs as systems evolve.

Thus, the EPOS solution overcomes the major problems associated with previous or parallel work namely: many-to-many interfaces between software brokers or systems and insufficiently rich metadata for automation while enabling interoperability across multiple asset sources.

On October the 30th 2018, the European Commission granted the legal status of European Research Infrastructure Consortium (ERIC) to EPOS, which was already promoted as a landmark in the ESFRI 2018 Roadmap.

The rest of the paper is organized as follows: Section II describes the architecture; Section III discusses the importance of metadata and Section IV discusses the major challenges faced currently and progress towards solutions and Section V gives the current state and outlook.

II. ARCHITECTURE

The ICT architecture of EPOS is designed to facilitate the research community and others in discovering and utilizing through the ICS the assets provided by the TCS communities.

A. Introduction

In order to provide end-users with homogeneous access to services and multidisciplinary data collected by monitoring infrastructures and experimental facilities (and to software, processing and visualization tools as well) a complex scalable and reliable architecture is required. A snapshot of the architecture is outlined in Figure 1. It includes three main layers:

Integrated Core Services – ICS, the core component designed and run by EPOS; this is the place where the integration of data and services provided by the TCS, Community Layer occurs. Integrated Core Services are characterized by a Central Hub (ICS-C), whose main goal is to host the metadata catalog and orchestrates external resources (e.g., HPC), and the Distributed Services (ICS-D), whose goal is to provide resources (e.g., computational, visualisation).

Thematic Core Services – TCS, made up of pan European e-Infrastructures, which disseminate data and services of a single discipline (e.g., seismology with ORFEUS/EIDA).

National Research Infrastructures – NRI, made up of RIs providing data and services,

Starting from the latter, NRI represent the wealth of assets provided by national or regional institutions or consortia, and are referred to as DDSS, i.e., Data, Data-products, Software and Services. The asset descriptions were collected first as DDSS in the DDSS master table (stored in Excel), which also records the state of maturity and management parameters. This is now being replaced progressively by the so-called Granularity Database

(GRDB), which records the same information as the DDSS master table but using the same metadata standard as that of the ICS-C catalog (described below in Section III) for ease of managing the process of approving a DDSS for inclusion in the ICS-C metadata catalog. The GRDB DDSS records are harvested as metadata for population of the EPOS ICS-C catalog.

TCSs enable the integration of data and services from specific scientific communities. The architecture of the services provided by the individual communities is not prescribed, what is required is that the metadata describing the data and services available is in a form that can be consumed by the ICS, allowing the ICS to integrate with those services and data (Figure 1).

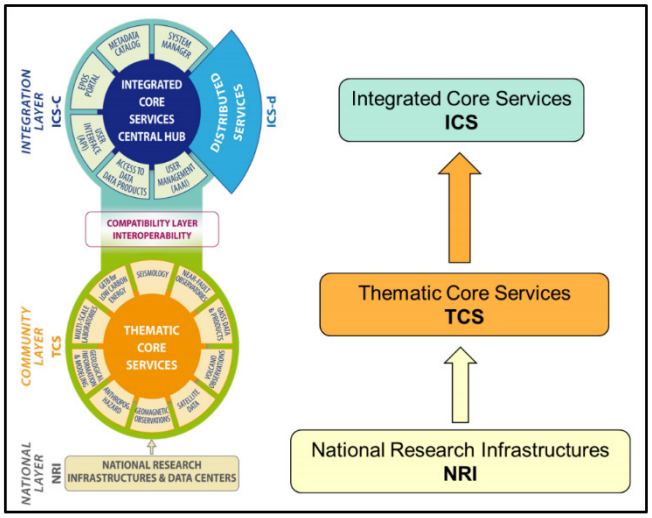


Figure 1. EPOS Architecture

B. ICS

The EPOS-ICS provides the entrypoint to the EPOS environment. The ICS consists of the ICS-C and distributed computational resources including also processing and visualisation services (ICS-D) of which a specialization is Computational Earth Science (CES). ICS-C provides a catalog of, and access to, the assets of the TCSs. It also provides access to e-Infrastructures (e-Is)as ICS-Ds upon which (parts of) workflows are deployed (other parts may be deployed within the computing capabilities of RIs within EPOS). EPOS has been involved in projects with e-Is to gain joint understanding of the interfaces and capabilities ready for deployment from ICS-C. EPOS has also been involved in the VRE4EICproject [18] (and cooperating with EVER-EST [19]) to ensure convergent evolution of the EPOS ICS-C user interface and APIs for programmatic access with the developing Virtual Research Environments (VREs). EPOS partners are also participating inthe recently approved ENVRIFAIR [20] project, which will assist in building linkages between EPOS ICS-C and European Open Science Cloud (EOSC) (Figure 2).

The linkage between ICS-C on the one hand and the e-Is and TCS local computing resources and assets on the other is through ICS-Ds, which will be constructed as a workflow in the ICS-C and managed in the deployment phase. The workflow for the deployment (which may be a simple file download or a complex set of services including analytics and visualisation) will be generated within the ICS-C by interaction with the users. The workflow will be checked by the end-user before deployment. However, the detailed content/capability of the assets might not be known, e.g., the dataset may not contain the relevant information despite its metadata description, or the software may not execute as the user expects despite the metadata description.

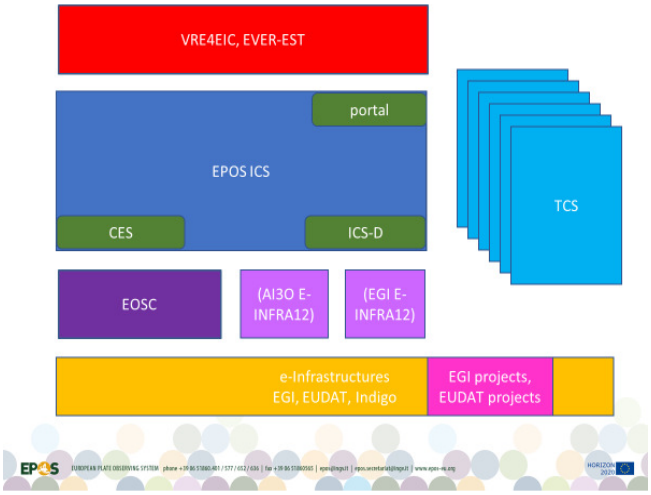


Figure 2. EPOS Positioning

The execution of the deployment is monitored and execution information is returned to the end-user. The workflow may be deployed in one of two ways: (a) directly with no user interaction during execution of the deployment; (b) step-by-step with user interaction (so-called computational steering) between each step. Deployments of type (a) will have better optimisation (for performance) and security but could possibly execute a workflow, the components of which do not behave as the user expects. Deployments of type (b) lack optimisation but allow the user to stop the workflow deployment at any step, examine the results and – if not as expected – reorganise the workflow (by changing components) to meet more closely the requirement.

The ICS represents the infrastructure consisting of services that will allow access to multidisciplinary resources provided by the TCS. These will include data

and data products as well as synthetic data from simulations, processing, and visualization tools.

C. ICS-C

The ICS-C consists of multiple logical areas of functionality, these include the Graphic User Interface (GUI), web-API, metadata catalogue, user management etc. A micro-service architecture has been adopted of the ICS-C, where each (micro) services is atomic and dedicated to a specific class of tasks. The ICS-C is where the integration of other services from ICS-D and TCS takes place. The architectural constraints for the ICS-D are elaborated as a metadata model within the ICS-C CERIF (Common European Research Information Format) [21] catalog and are being implemented.

The ICS-C System is the main system that manages the integration of DDSS from the communities. On top of such a system, a Graphic User Interface (GUI) enables the user to search, discover and integrate data in a user-friendly way.

The EPOS ICS-C system architecture (Figure 3) was designed and developed with the aim of integrating data and services provided by TCS. In order to a) enable the system to run in a distributed environment, b) guarantee up-to-date technological upgrades by adopting a software-independent approach, c) proper scaling of specific system functionalities, the chosen architecture followed a microservices paradigm.

The Microservices architecture approach envisages small atomic services dedicated to the execution of a specific class of tasks, which have high reliability [22], [23]. Such architecture replaces the monolith with a distributed system of lightweight, narrowly focused, independent services. In order to implement the microservices paradigm, Docker Containers technology was used [24]. It enables complete isolation of independent software applications running in a shared environment. In particular, each microservice is developed in the Java language and performs a simple task, as atomic as possible. The communication between microservices is done via messages received and sent on a queueing system, in this case RabbitMQ [25]. As a result, a chain of microservices processes the requests.

The current architecture includes an Authentication, Authorisation, Accounting Infrastructure (AAAI) module. This has been implemented using UNITY [26] and has involved close cooperation with CYFRONET. Since May 2018 this has formed the basis of anintegrated authentication system for academic communities Authorisation is more complex and depends on rules agreed with the TCS (within the context of the financial, legal and governance traversal workpackages of EPOS-IP) for each of their assets and included further metadata elements into the CERIF catalog to control such authorisation. AAAI will be continuously evolved and updated to ensure appropriate security, privacy and governance. Related to this, the GUI now provides a user notification pointing to a legal disclaimer for the EPOS system, terms and conditions and acceptance of cookies.

A major requirement of the system, after asset discovery, is the construction of workflows that can be used to access / process data. This has implications for the entire software stack; visually designing the workflows, managing and persisting inputs and outputs, scheduling and execution of processes, access to metadata, access to data and service from the TCS. The topic as whole required significant analysis of requirements and available technologies. Working in cooperation with the VRE4EIC project we have the basic components for (a) a general workflow manager interface; (b) interfaces to specific workflow managers such as Taverna [2].

Beyond simple map visualisations that consume web map services the ICS-C user interface may be required to support additional types of visualisation. This set of supported visualisation types and associated data formats is being confirmed with the TCS representatives through a series of ongoing workshops as it will not be practical to support all formats of data for all types of visualisation.

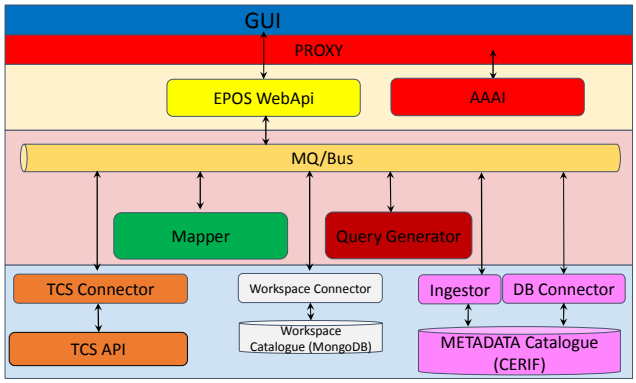


Figure 3. ICS-C Architecture

D. ICS-D

The distributed services offered by the ICS-D facet of the architecture ties-in with the workflow management, as the distributed services in question - beyond just being discoverable - are likely candidates for inclusion in processing workflows. A specification of the metadata elements required for ICS-D has been produced, is under review and forms part of the architecture. ICS-D will appear to the workflow, or to the end-user, as a service accessed through an API. However, the choice of which ICS-D to use and the deployment of a workflow across one or more ICS-Ds requires optimisation middleware.

Results from the PaaSage project [27] are relevant and the concurrent MELODIC project [28] offers optimisation including that based on dataset placement and latency. Further refinement of requirements and the architectural interfaces continues.

III. METADATA

Metadata is the key to discover and utilise the heterogeneous assets of EPOS in a homogeneous way thus facilitating cross-domain, interoperable science.

A. Introduction

The metadata catalogue is the key technology that enables the system to manage and orchestrate all resources required to satisfy a user request. By using metadata, the ICS-C can discover data or other digital objects requested by a user, contextualise them (for relevance and quality) access them, send them to a processing facility (or move the code to facility holding the data) depending on the constructed workflow, and perform other tasks. The catalogue contains: (i) technical specification to enable autonomic ICS access to TCS discovery and access services, (ii) metadata associated with the digital object with direct link to it, (iii) information about users, resources, software, and services other than data services (e.g., rock mechanics, geochemical analysis, visualization, processing). The data model used for the catalogue is CERIF.

Metadata describing the TCS DDSS are stored using the CERIF data model, which differs from most metadata standards in that it (1) separates base entities from linking entities thus providing a fully connected graph structure; (2) using the same syntax, stores the semantics associated with values of attributes both for base entities (to ensure valid attribute values are recorded for instances, e.g., ISO country codes) and for linking entities (for role of the relationship), which also store the temporal duration of the validity of the linkage. This provides great power and flexibility. CERIF also (as a superset) can interoperate with widely adopted metadata formats such as DC (Dublin Core) [29], DCAT (Data Catalogue Vocabulary) [30], CKAN (Comprehensive Knowledge Archive Framework) [31], INSPIRE (the EC version of ISO 19115 for geospatial data) [32] and others using convertors developed as required to meet the metadata mappings achieved between each of the above standards and CERIF. The metadata catalogue also manages the semantics, in order to provide the meaning of the instance attribute values. The structure of base entities and linking entities used for metadata instances is also used for the semantic layer of CERIF; the base entities containing lexical entries and the linking entities maintaining the relationships between them allowing a full ontology graph structure including not only subset and superset terms but also equivalent terms (especially useful for multilinguality) and any other role-based logical relationship between terms.

The use of CERIF provides automatically:

(a) The ability for discovery, contextualization and (re-)use of assets according to the FAIR principles [33];
(b) A clear separation of base entities (things) from link entities (relationships);
(c) Formal syntax and declared semantics;
(d) A semantic layer also with the base/link structure allowing crosswalks between semantic terminology spaces;
(e) Conversion to/from other common metadata formats;
(f) Built-in provenance information because of the timestamped role-based links;
(g) Curation facilities because of being able to manage versions, replicates and partitions of digital objects using the base/link structure.

The catalog is constantly evolving with the addition of new assets (such as services, datasets) but also increasingly rich metadata as the TCSs improve their metadata collection to enable more autonomic processing.

B. TCS Metadata

The process of populating the catalog is crucial in the EPOS vision. Indeed, populating the catalog means to make available all the information needed by an end user to perform queries, data integration, visualisation and other functionalities provided by the system.

Greater interaction with TCS communities to ensure that their metadata, data and services are available for harvesting in the appropriate format and to populate the CERIF data model has been achieved and will be continued.

C. ICS Metadata

In order to manage all the information needed to satisfy user requests, all metadata describing the TCS Data, Datasets, Software and Services is stored into the EPOS ICS, internal catalog, based on the aforementioned CERIF model, which differs from most metadata standards used by various scientific communities in that it is much richer in syntax (structure) and semantics (meaning).

For this reason, EPOS ICS has sought to communicate to the TCS communities the core elements of metadata required to facilitate the ICS through the EPOS Metadata Baseline. This baseline can be considered as an intermediate layer that facilitates the conversion from the community metadata standards such as ISO19115/19, DCAT, Dublin Core, INSPIRE etc. describing the DDSS

elements and not the index or detailed scientific data (Figure 4).

The EPOS baseline presents a minimum set of common metadata elements required to operate the ICS taking into consideration the heterogeneity of the many TCSs involved in EPOS. It has been implemented as an application profile using an extension of the DCAT standard, namely the EPOS-DCAT-AP [34]. It is possible to extend this baseline to accommodate extra metadata elements where it is deemed that those metadata elements are critical in describing and delivering the data services for any given community. Indeed, this has happened when the original EPOS-DCAT-AP was found to be inadequate and a new version with richer metadata was designed and implemented.

The metadata to be obtained from the EPOS TCSs as described in the baseline document (and any other agreed elements) are mapped to the EPOS ICS CERIF catalog. The process of converting metadata acquired from the EPOS TCS to CERIF isdone in consultation with each TCS as to what metadata they have available and harvesting mechanisms.

The various TCS nodes have APIs or other mechanisms to expose the metadata describing the available DDSS in a TCS specific metadata standard that contains the elements outlined in the EPOS baseline documents better described in the following sections. It also requires ICS APIs (wrappers) to map and store this in the ICS metadata catalogue, CERIF. These APIs and the corresponding ICS convertors collectively form the "interoperability layer" in EPOS, which is the link between the TCSs and the ICS.

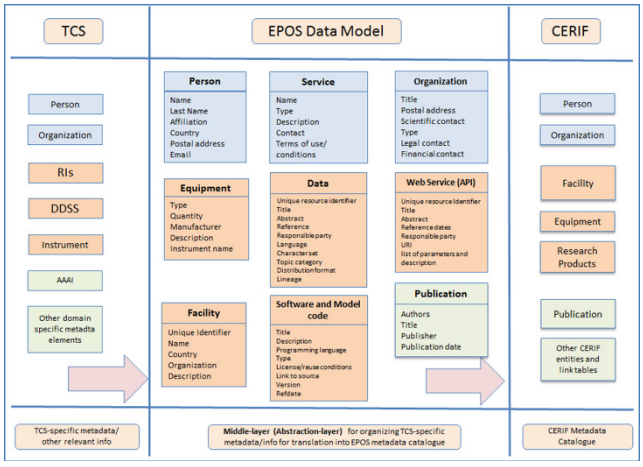


Figure 4. EPOS Metadata Baseline

D. DDSS and Granularity Database

As a part of the requirements and use cases collection (RUC) from the TCSs, a specific list was prepared to include all data, data products, software and services (DDSS). This DDSS Master Table was used as a mechanism to update the RUC information as well as providing a mechanism for accessing more detailed IT technical information for the development of the ICS Central Hub (ICS-C). The DDSS Master Table was also used for extracting the level of maturity of the various DDSS elements in each TCS as well as providing a summary of the status of the TCS preparations for the ICS integration and interoperability. The current version of the DDSS Master Table consists of 368 DDSS elements, where 201 of these already exist and are declared by TCSs to be ready for implementation. The remaining DDSS elements required more time to harmonize the internal standards, prepare an adequate metadata structure and so are available for implementation soon. In total, 21 different harmonization groups (HGs) are established within the EPOS-IP project to help organizing the harmonization issues in a structured way. TCSs are preparing individual TCS Roadmaps, which will describe the development and implementation plans of the remaining DDSS elements including a time-line and resource allocations. In addition, user feedback groups (UFGs) are being established in order to give constant and structured feedback during the implementation process of the TCS-ICS integration and the development of the ICS.

The DDSS Master Table was constantly being updated as new information from the TCS WPs arrive. The older versions are also kept in the archive for future reference. The DDSS master table is being transformed to the GRDB (granularity database) because of the problems of referential and functional integrity using a spreadsheet; relational technology provides appropriate constraints to ensure integrity. As such, the GRDB represents a structured way of requirements and use cases collection (RUC) from the TCS communities. Updates or new entries to GRDB can be done either using a dedicated GUI or in an automated manner.

The TCS requirements and use cases (RUC) collection process was designed carefully, taking into account the amount and complexity of the information involved in all 10 different TCSs. An increasingly detailed RUC collection process is formulated and explained through dedicated guidelines and interview templates. A roadmap for the ICS-TCS interactions for the RUC collection process was prepared for this purpose and distributed to all TCSs.

In this approach, a five-step procedure is applied involving the following:

- Step 1: First round of RUC collection for mapping the TCS assets;
- Step 2: Second round of RUC collection for identifying TCS priorities;
- Step 3: ICS-TCS Integration Workshop for building a common understanding for metadata;
- Step 4: Third round of RUC collection for refined descriptions before implementation;
- Step 5: Implementation of RUC to the CERIF metadata.

Planning for the requirements and use cases (RUC) elicitation process started with the pre-project meeting held during the period July 8-9 2015 at the BGS (British Geological Survey) facilities in Nottingham, UK. The first version of the guidelines level-1 for the ICS-TCS integration was prepared soon after this meeting and was distributed to the TCS leaders and the relevant IT-contacts. A second, more detailed guidelines level-2 was prepared in September 2015 and distributed in the EPOS-IP project kick-off meeting held in Rome, Italy, during the period October 5-7 2015. Prior to the kick-off meeting, a preliminary collection of the RUC was requested from each TCS, which was then presented during the meeting.

In parallel with the guidelines for the ICS-TCS Integration, a dedicated RUC interview template level-1 was preparedto be used during the first site visits to the TCSs. The site visits were conducted during the time period between November 2015 and March 2016. All four steps are now completed, whereas step 5 with metadata implementation has started in January 2017 and is ongoing.

Work is almost complete in converting the DDSS tables (in Excel) to the GRDB using Postgres. This will (a) facilitate finding particular DDSS elements, eliminating duplicates and checking the progress of getting DDSS elements into metadata format; (b) actually harvesting to the metadata catalog.

IV. CURRENT CHALLENGES

This section lists the current challenges being addressed, beyond the system as described in [1].

A. Introduction

A project as large in terms of organisations, persons and assets involved and as complex in terms of governance, funding and technology required, necessarily faced many challenges. Some of the key challenges are discussed.

B. Metadata Conversion

As discussed in Section III, the use of a canonical rich metadata format is key to providing homogeneous access to the heterogeneous assets within EPOS. Reaching the state of all assets recorded in this standard posed some challenges. These are outlined below.

1) Heterogeneity

However, the multiple metadata ‘standards’ used widely within the various EPOS communities – and in some cases used by those communities within an international context for exchange of data – needed to be respected while converting to the canonical rich metadata standard CERIF. This conversion was achieved by much discussion between each TCS community and the ICS ICT team. The discussion involved understanding not only the metadata model being used (which usually was well-documented) but also how it was used – with which interpretation of the ‘rules’ of the model. As well as the heterogeneity in the ‘standards’ used, there was also heterogeneity in its interpretation, even of the same ‘standard’.

2) Complexity

CERIF provides a rich metadata model. Mathematically it is a fully connected graph. The metadata ‘standards’ used by the TCS communities were – in general – simple, consisting of records not unlike a library catalog card with attributes related to an asset such as a service or dataset. These attributes commonly included persons and organisations, which could be multiple and were not functionally dependent on the asset being described; this meant that the TCS metadata records did not have referential and functional integrity. However, the TCS communities were familiar with their own ‘standard’ and found difficulty in understanding (a) the concept of integrity to ensure validity of the metadata; (b) the need for a fully connected graph structure to represent more accurately the real world. As described in Section III, this problem was overcome by using a simplified intermediate format (EPOS-DCAT-AP), which – stored in RDF (Resource Description Framework) - acted as a ‘bridge’ between the simple metadata structures of the TCSs and the richness of CERIF.

C. Legal, Governance and AAAI Aspects

The overall intention of EPOS is to make assets findable, accessible, interoperable and reusable in an open environment and toll-free to not-for-profit users. However, it was necessary to introduce some technical ICT features to accommodate legal, governance and AAAI aspects.

1) Terms and Conditions of Use

A conditions of use document was produced and made accessible from the ‘landing page’ (the screen first encountered when accessing EPOS) with a requirement that a user should accept the Terms and Conditions.

2) Disclaimer

Similarly, a disclaimer document was produced and made accessible from the 'landing page' with a requirement that a user should accept the Terms and Conditions.

3) Cookies

Also, on the 'landing page' there is a requirement for the user to accept (or not) the use of cookies in EPOS.

4) AAAI – Authentication

There is a need to authenticate users (i.e., ensure the user has credentials to assure that they are who they claim to be) for several reasons. (a) it provides security against individuals accessing the system with malicious intent; (b) it allows audit and provenance trails to be related to a person for several purposes: to provide records to demonstrate compliance with GDPR (General Data Protection Regulation); to allow reproduction of the scientific pathway to corroborate research results; to improve user interaction by suggesting (based on past usage) assets to be used. EPOS aligns with current leading-edge work in this area using authentication agents such as EduGAIN [35] and also tracks the ongoing work within the European AARC2 project [36].

5) AAAI-Authorisation

Once a user is authenticated, he/she may be authorized (by some other authority) to access assets. The access may be restricted by role of the user, by time interval, by the process intended (e.g., read, execute, modify, delete) as well as by collection of assets or individual asset. The authorization system is currently being discussed with the TCS representatives since (a) it requires collection of more metadata for the assets, persons and organisations; (b) it requires appropriate access control program code to be provided.

D. Use of DoI (Digital Object Identifier)

A problem for a particular collection of assets is the use of DoI. The DoI system works by dereferencing the DoI to a landing page, which contains text describing the asset and a URL, which dereferences to the asset itself. The concept is based on human interaction, the human reads the landing page text and decides whether to access the asset.

In contrast, the EPOS ICS-C is based around the concept that the user queries the metadata catalog for assets that – satisfying the query - are relevant and of sufficient qualityto allow automated access - and then accesses them directly.

Two solutions are being worked upon: (a) for those DoI-based collections, which have a well-structured landing page template to use MIME types to access the URL pointing directly to the asset, thus 'bypassing' the step of a human reading the landing page (although the lack of rich metadata in the metadata catalog may well mean that relevant assets are not recalled by the query); (b) where the landing page text is well-structured, converting the metadata text of the landing page to a CERIF record in the metadata catalog together with the asset URL thus rendering the landing page redundant.

E. Complexity of the GUI (Graphical User Interface)

Different TCSs have different ways of finding, accessing, interoperating and re-using assets. The design challenge was to find a common process structure with step sequences (including cycling back to previous steps) to accommodate these different requirements. In turn this made the design of the GUI more complex since different users wished to traverse the process steps in different ways. At workshops involving TCS community representatives and the ICS ICT team scientific stories were mapped to use cases, and these were used to define the GUI requirements.

The complexity arises because users may wish to confirm their choice of a single asset by seeing it visually – on a map or chart – before deciding whether to add the metadata for that asset to the workspace that they are constructing during the session. Furthermore, they may wish to change the parameters of the asset – especially of a service – and re-visualise. On the other hand, some users wish to see the assets represented by metadata in the workspace visualized as a 'build-up' with each one overlaid on the other. Thereafter they may wish to change the parameters of one or more assets before composing a workflow, which involves cycling back to visualization of single assets before checking again the 'build-up' of visualisations for all the assets represented by metadata in the workspace.

Different possibilities are being tested with representative TCS users to determine which options should be implemented in the operational system due to be released end-September 2019.

F. Intersection of AAAI with GUI

Another challenging factor is the question of when to demand that a user is authenticated. Some (few) users wish completely anonymous, open, toll-free access. This is clearly not possible for legal and governance reasons; for example, the potential for liability litigation or the potential use of a large amount of supercomputing resources without prior authorization.

There is, however, an argument for a user being able to query the metadata catalog and visualize individual selected assets to see if they suit his/her requirements

before logging in / authenticating prior to composing or deploying a workflow. Furthermore, this approach leaves the TCS communities free to control authorization of asset usage since login and authentication takes place before access to the assets with authorization. However, this approach leaves metadata catalog access open to a liability challenge (since the user will not yet have accepted the disclaimer) and also may contravene GDPR (since the metadata includes information about persons - such as the owner of an asset or the manager of an asset).

The safest approach is to demand login / authentication at session start. This ensures not only security and legal/governance compliance but also initiates appropriate audit and provenance recording. The counter-argument is that immediate login/authentication may be a barrier to use of the system for some users.The ICS ICT team is currently discussing these options with both EPOS governance structures and TCS users to find the appropriate design that can be implemented.

V. CONCLUSION AND FUTURE WORK

The European plate observing system (EPOS) is addressing the challenges of accessing heterogeneity in a homogenous way by building an integration node called Integrated Core Services. This system is metadata driven and uses the CERIF model. Currently 136 distinct DDSS accessible through 264 different web-services from the domain communities are represented by CERIF metadata in the EPOS ICS-C catalog. These services, described by the metadata, can be discovered, contextualised and utilised individually or composed into workflows and hence become interoperable. A GUI (Graphical User Interface) provides the user view onto the catalog, and it also provides a workspace to collect the metadata of the assets selected for use (Figure 5). From the workspace a workflow may be constructed and deployed.

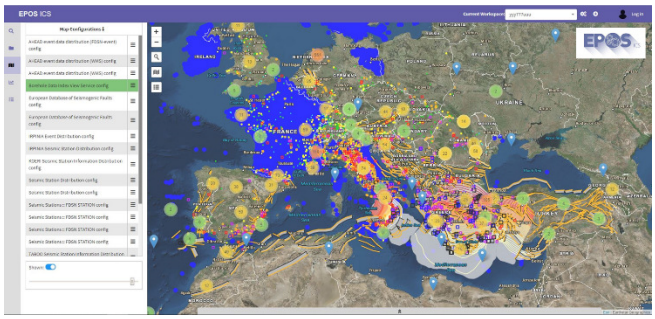


Figure 5. EPOS-ICS graphical user interface.

Future plans include:

a) Harvesting of metadata describing more assets: not only services but also datasets, software, workflows, equipment;
b) Improving the GUI to allow workflow deployment with 'fire and forget' technology or single-step with user checking and adjustment at each step;
c) Completion of the (current prototype) software to permit trans-national access to laboratory and sensor equipment;
d) Improved AAAI (Authentication, authorisation, accounting infrastructure) to give the domain users finer-grained control over access to their assets;
e) The inclusion of virtual laboratory-type interfaces (virtual research environments) allowing users access and connectivity including open-source frameworks such as Jupyter notebooks [3], which are increasingly being used in some scientific communities.

The architecture outlined and demonstrated (in successive prototypes) in EPOS-IP has found favour (not without some criticism of course – leading to agile improvements) from the user community. Furthermore, the prototype system has passed Technological Readiness Assessment procedures within the governance of the EPOS-IP project. Currently the ICS is undergoing validation tests. The first operational release is scheduled for end-September 2019. The architecture meets the requirements, it is state of the art and has a further development plan.

ACKNOWLEDGMENT

The authors acknowledge the work of the whole ICT team in EPOS reported here and the funding of the European Commission H2020 program (Grant agreement 676564) and National Funding Councils that have made this work possible.

REFERENCES

[1] K. Jeffery, D. Bailo, K. Atakan, and M. Harrison, "EPOS: European Plate Observing System," in Proc. Eleventh International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2019), pp. 79-86.
[2] Taverna: https://taverna.incubator.apache.org/2019.11.11
[3] Jupyter: https://jupyter.org/2019.11.11
[4] F. Pérez and B. Granger, "IPython: a system for interactive scientific computing," Computing in Science and Engineering, 9(3), pp. 21-29, June 2007.
[5] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, and P. Ivanov, "Jupyter Notebooks-a publishing format for reproducible computational workflows," in Proc. 20th International Conference on Electronic Publishing (ELPUB), pp. 87-90, May 2016.

[6] PRACE: http://www.prace-ri.eu/ 2019.11.11
[7] EOSC pilot: https://eoscpilot.eu/ 2019.11.11
[8] GEANT: http://www.geant.org/ 2019.11.11
[9] EGI: https://www.egi.eu/ 2019.11.11
[10] EUDAT: https://eudat.eu/ 2019.11.11
[11] OpenAIRE: https://www.openaire.eu/ 2019.11.11
[12] P. Sutterlin, K. Jeffery, and E. Gill, "Filematch: A Format for the Interchange of Computer-Based Files of Structured Data," Computers and Geosciences, Vol. 3 (1977), pp. 429-468.
[13] UMM: https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm 2019.11.11
[14] Geonetworkhttps://geonetwork-opensource.org/ 2019.11.11
[15] EarthCube: https://www.earthcube.org/ 2019.11.11
[16] AuScope: http://www.auscope.org.au/ 2019.11.11
[17] GEOSS: https://www.earthobservations.org/geoss.php 2019.11.11
[18] VRE4EIC: https://www.vre4eic.eu/ 2019.11.11
[19] EVEREST: https://ever-est.eu/ 2019.11.11
[20] ENVRI-FAIR website http://envri.eu/envri-fair/ 2019.11.11
[21] CERIF: https://www.eurocris.org/cerif/main-features-cerif 2019.11.11
[22] S. Newman, "Building Microservices," O'Reilly Media, Inc., February 2015, ISBN: 9781491950340.
[23] D. Namiot and M. Sneps-Sneppe, "On Microservices Architecture," International Journal of Open Information Technologies, ISSN 2307-8162, Vol. 2, No. 9, pp. 24-27, 2014.
[24] Docker: https://www.docker.com/ 2019.11.11
[25] RabbitMQ: https://www.rabbitmq.com/ 2019.11.11
[26] UNITY: http://www.unity-idm.eu 2019.11.11
[27] PaaSage: https://paasage.ercim.eu/ 2019.11.11
[28] MELODIC: melodic.cloud/ 2019.11.11
[29] DC: http://dublincore.org/documents/dces/ 2019.11.11
[30] DCAT: https://www.w3.org/TR/vocab-dcat/ 2019.11.11
[31] CKAN: https://ckan.org/ 2019.11.11
[32] INSPIRE: https://inspire.ec.europa.eu/ 2019.11.11
[33] FAIR: https://www.force11.org/grohttps://ckan.org/up/fairgroup/fairprinciples 2019.11.11
[34] EPOS-DCAT-AP on GitHub: https://github.com/epos-eu/EPOS-DCAT-AP 2019.11.11
[35] https://edugain.org/ 2019.11.11
[36] https://aarc-project.eu/ 2019.11.11

Towards an Automated Printed Circuit Board Generation Concept for Embedded Systems

Tobias Scheipel and Marcel Baunach

Institute of Technical Informatics
Graz University of Technology
Graz, Austria
E-mail: {tobias.scheipel, baunach}@tugraz.at

***Abstract*—Future embedded systems will need to be generic, reusable and automatically adaptable for the rapid advance development of a multitude of different scenarios. Such systems must be versatile regarding the interfacing of electronic components, sensors, actuators, and communication networks. Both the software and the hardware might undergo a certain evolution during the development process of each system, and will significantly change between projects and use cases. Requirements on future embedded systems thus demand revolutionary changes in the development process. Today these processes start with the hardware development (bottom-up). In the future, it shall be possible to only develop application software and generate all lower layers of the system automatically (top-down). To enable automatic Printed Circuit Board (PCB) generation, the present work deals mainly with the question *"How to automatically generate the hardware platform of an embedded system from its application software?"*. To tackle this question, we propose an approach termed *papageno*PCB, which is a part of a holistic approach known as *papageno*X. This approach provides a way to automatically generate schematics and layouts for printed circuit boards using an intermediate system description format. Hence, a system description shall form the output of application software analysis and can be used to automatically generate the schematics and board layouts based on predefined hardware modules and connection interfaces. To be able to edit and reuse the plans after the generation process, a file format for common electronic design automation applications, based on Extensible Markup Language (XML), was used to provide the final output files.**

Keywords–embedded systems; printed circuit board; design automation; hardware/software codesign; systems engineering.

I. INTRODUCTION

Embedded systems are of relevance in virtually every area of our society. From the simple electronics in dishwashers to the highly complex electronic control units in modern and autonomous cars – daily life today is nearly inconceivable without those systems. As the technology improves, the complexity of embedded systems inevitably and steadily increases. A whole team of engineers usually plans, designs, and implements a novel system in several iteration steps. An example of such a process in the automotive industry is shown in Figure 1.

Designing an embedded system can be prone to errors due to a multitude of possible error sources. This presents one major challenge when designing such a system: The challenge of how to eliminate error sources and make design processes more reliable and, therefore, cheaper. Most design paradigms today choose a bottom-up approach. This means that a suitable computing platform is chosen after defining all requirements with respect to these explicit requirements, prior experience, or educated guesses. Then, software development can either start based on an application kit of a computing platform, or some prototyping hardware must be built beforehand. If the requirements change during the development process, major problems could possibly arise, e.g., new software features cannot be implemented due to computing power restrictions or additional devices cannot be interfaced because of hardware limitations. Another problem could arise if connection interfaces or buses become overloaded with too much communication traffic after the hardware has already been manufactured.

To tackle these problems, we already proposed a holistic approach, ***papageno***X, and a sub-approach, ***papageno***PCB in [1]. In the course of this work, we intend to further discuss and extend our approach in more detail in the following sections. ***papageno***X is a novel approach that has been developed for use while creating embedded systems with a top-down view. Therefore, it uses application source code to automatically generate the whole embedded system in hardware and software. One part of the concept behind this approach is ***papageno***PCB. This concept handles the automatic generation of schematics and board layouts for PCB design with standardized XML-based [2] output from intermediate system description models. To do so, a module-based description of the system hardware and software needs to be made. Furthermore, connections between the hardware modules on wire level are done automatically. The concept and its related challenges were the main topics of the work described in this paper, whereas software analysis and model generation is part of work that will be conducted in the future with ***papageno***X.

The paper is organized as follows: Section II includes a summary of related work. The rough idea of the holistic vision of ***papageno***X (this paper includes a detailed description of the first part of this concept) is illustrated in Section III, whereas Section IV starts with the system description format within ***papageno***PCB. In Section V, an explanation is given of the necessary steps taken to create the final output, and Section VI includes a proof of concept example, an analysis of the scalability and performance for the developed generator and a use case with a manufactured prototype. The paper concludes with Section VII, in which the steps that need to be taken to achieve a final version of ***papageno***X are described.

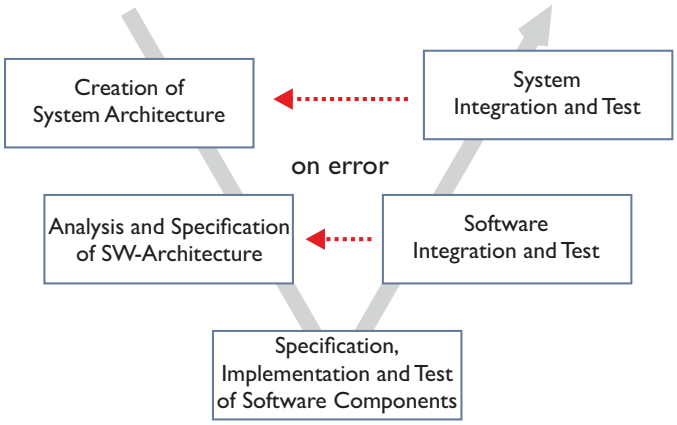


Figure 1. Automotive design process according to the V-Model [3].

II. State of the Art and Related Work

This section gives an overview on how embedded systems are developed nowadays and how hardware can be generated automatically in different types of systems. Additionally, some approaches towards software annotations and design space exploration are shown to provide an overview.

A. Embedded Systems Prototyping Approaches

Conventional embedded systems prototyping makes use of very specialized hardware platforms, capable of executing a vast variety of use cases typical for the field of deployment (e.g., an automotive Electronic Control Unit (ECU), a Cyber-Physical System (CPS), an Internet-of-Things (IoT) device).

In the context of ECU prototyping platforms, one approach is *rCube2* [4], based on two powerful independent TC1797 [5] Microcontroller Units (MCUs). The two processors can interact via shared memory, but are completely isolated during execution. AVL RPEMS [6] is a generic engine control platform provided as a highly flexible and configurable engine management system for the development and optimization of conventional and new combustion engines, power and emissions optimization, and the realization of hybrid and electric powertrains. The current version consists of a single-core automotive MCU (TC1796 [7] or TC1798 [8]) with different variants for diesel and gasoline engine control applications. These different PCBs are equipped with automotive-compliant Application-specific Integrated Circuits (ASICs) and a head-mounted MCU board, which allows prototyping as close as possible to series production. It was designed to offer engineers utmost flexibility when developing new control algorithms for non-standard engines, or standard engines with new components.

There are other similar prototyping platforms from commercial suppliers, but they also lack in flexibility and adaptability when it comes to hardware changes. The main problem of those commercial solutions is that even though they offer high performance and come with complete toolchains, their hardware is very different from a series device, as overcompensation takes place. Since the components cannot be easily changed when a prototype is turned into a commercial product, a complete redesign has to take place.

When the need for hardware changes after deployment is taken into account, reconfigurable logic is mostly mentioned in literature. This can reach from pure Field Programmable Gate Arrays (FPGAs) to System on Chips (SoCs), which include an FPGA alongside other MCU cores (e.g., Zynq-7000 [9]). The main advantage of those systems is that one does not have to change the physical hardware, but can easily adapt features like on-chip peripherals within the logic without the need of manufacturing an ASIC. However, it is not possible to change physical hardware features after deployment with those devices.

B. Automatic Hardware Generation

As this work is concerned with the automatic generation of hardware and extensively utilized hardware definition models, it was influenced by existing solutions such as devicetree, which is used, e.g., within Linux [10]. The devicetree data structure is used by the target Operating System's (OS) kernel to handle hardware components. The handled components can comprise processors and memories, but also the internal or external buses and peripherals of the system. As the data structure is a description of the overall system, it must be created manually and cannot be generated in a modular way. It is mostly used with SoCs and enables the usage of one compiled OS kernel with several hardware configurations. As far as automatic generation of schematics from software is concerned (top-down), there are a few solutions towards design automation. Some papers deal with the question, how to generate schematics, so that these look nice for a human reader by using expert systems [11]. Some work on the generation of circuit schematics has even been done by extracting connectivity data from net lists [12]. These approaches all have some kind of network information as a basis and do not extract system data out of – or are even aware of – application source code or system descriptions.

C. Annotations and Design Space Exploration

Different approaches have been taken to use annotated source code to extract information about the underlying system. Annotations can be used to analyze the worst-case execution times [13][14] of software in embedded systems. Other approaches that have been taken have used back-annotations to optimize the power consumption simulation [15]. These annotations have allowed researchers to gain a better idea of how the system works in a real-world application, meaning that the annotated information is based on estimations or measurements. Introduction of annotations can be achieved by simple source code analysis or more sophisticated approaches such as, e.g., creating add-ons or introducing new features into source code compilers.

To generate systems out of application software, annotations can be used to extract requirements. These requirements can then be utilized to apply design space exploration [16] by, e.g., modeling constraints [17]. In [18], different types for design space explorations are shown and categorized, also mentioning language-based constraint solvers featuring, e.g., MiniZinc [19]. By using approaches like these, a design space model can easily be translated into a mathematical model for optimization.

All the approaches and concepts mentioned above have some advantages and inspired this work, as no solution has yet been proposed for how to automatically generate PCBs from source code.

III. Main Idea of *papageno*X

The main idea of ***papageno***X consists of an application driven electronics generation and the inversion of the state of the art "software follows or adjusts to hardware" paradigm in embedded systems development, where the design starts with the hardware architecture. Software is then built on selected components (e.g., automotive grade MCUs and PCBs).

Even when hardware deficits become visible during the software development process, the hardware is unlikely to see significant changes due to the high cost and many people or even companies involved. Thus, software developers try to compensate, e.g., by manual tuning and workarounds beyond automotive standards (e.g., AUTOSAR [20]). This violates compliance and is one reason why prototypes differ significantly from series devices, also complicating the transition and the subsequent maintenance in the field. Apart from this, future embedded systems will contain reconfigurable logic which is scarcely supported in current development processes due to both the lack of support and a fear of even more complexity (in addition to the software, electronics, and networks). This is why ***papageno***X is an abbreviation for **P**rototyping **AP**plication-based with **A**utomatic **GEN**eration **O**f **X**. The envisioned concept of it will prospectively contain a set of tools that can be used to automatically generate the software, reconfigurable logic, and hardware of the final prototype of system X by simply using application software source code. In this context, system X could be an automotive ECU, a CPS, an IoT device or some other embedded system. The goal is to support frequent changes to the Application Software (ASW) requirements by immediately reflecting them in the Basic Software (BSW), logic, and electronics – reducing time to market and efforts in development and maintenance. During development, the process will optimize the selection and configuration of BSW, on-board components, network interfaces, etc. for simplified transition to series production ("perfect fit"). After deployment, the process will help in the assessment of intended ASW chances to quantify the consequences on lower layers and thus to evaluate their feasibility and cost.

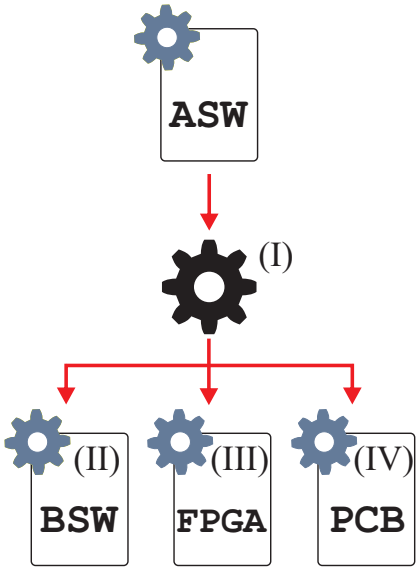


Figure 2. The main idea behind the ***papageno***X approach.

As depicted in Figure 2, the starting point of ***papageno***X is some application as a model or in source code. This ASW is analyzed in order to get to know all necessary requirements for the underlying system layers. These requirements are then used to generate software code that includes BSW and an executable ASW, reconfigurable logic code in a hardware description language (e.g., VHDL [21], Verilog [22]) for FPGAs, as well as schematics and layouts for PCBs. In this context, the term BSW subsumes operating systems with, e.g., drivers, services, hardware abstraction. Even though the ***papageno***X approach envisions the generation of reconfigurable logic, it differs from, e.g., SystemC [23], because it also generates hardware on the PCB level.

The following steps are envisioned within ***papageno***X:

1) application software development
2) in-depth analysis of ASW with respect to functional and non-functional requirements (NFRs)
3) creation of a selection space over potential components
4) filtering of the selection space with respect to general design decisions (e.g., data retention time)
5) generation of potential configurations from components
6) evaluation and optimization towards NFRs to select a single or several final, best fitting configuration(s)
7) mapping of functions or algorithms to reconfigurable logic (FPGAs)

To get a simple overview, the following example sketches the envisioned process while developing an embedded system with our novel approach: a user wants to store data somewhere permanently; with a data rate $\geq 5MB/s$ by writing this line of code in the ASW:

```
store_data(&data, StoreType.Permanent, 5000000);
```

The follow-up analysis of the ASW yields in an exemplary selection space as depicted in Figure 3. The green filled boxes illustrate the final configuration selected by the concept.

So, apart from the running application on the topmost level, a BSW must be generated, supporting a FAT16 [24] file system on top of a SD card driver and its underlying Serial Peripheral Interface (SPI, [25]) module driver. But even more important for this work, the final embedded system must be composed of a computing platform and a storage device, interconnected with each other. Finally, the generated system structure must be manufactured on a PCB, still matching all requirements with its properties.

Based on this overview, ***papageno***X will contain four major parts (also depicted in Figure 2):

(I) ASW analysis → creates selection space
(II) BSW generation → derives, e.g., needed components, drivers, OS features
(III) FPGA generation → maps functions to reconfigurable logic
(IV) PCB generation → module-based generation of suitable PCBs

In this paper, however, the main focus is on (IV), where a very first step is taken to generate a PCB from an intermediate system model (prospectively extracted from source code). The attempt is made to answer the research questions "What information is needed to automatically generate PCBs from ASW?" and "How can this information be used to generate a PCB prototype matching all ASW requirements?". It is henceforth named ***papageno***PCB.

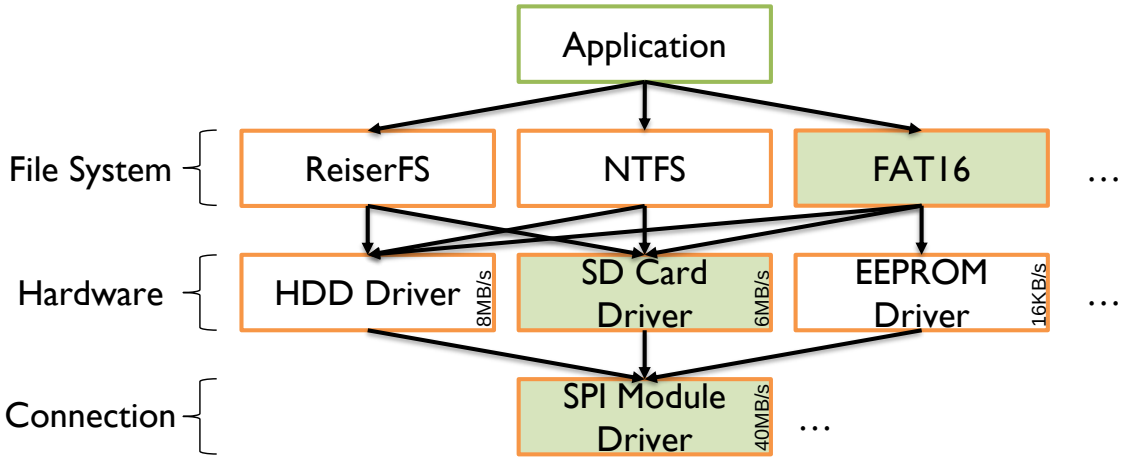


Figure 3. A system's requirements mapped to a corresponding exemplary selection space.

IV. System Description Format

The system description format in ***papageno***PCB is module-based. This means that every possible module, e.g., a MCU board or different peripherals must be defined before they are connected with each other. The whole description and modeling approach taken is generic, which enables its easy adaptation to different use cases. The structure was defined according to a JavaScript Object Notation (JSON) [26] format, and three different kinds of definition files were established:

A. Module Definition: One single file that defines the hardware module, its interfaces and its pins, and a second file that contains the design block for creating schematics and board layouts concerning this module.

B. Interface Definition: Generic definition of several different interface types to interconnect modules with each other; new types can be easily implemented and included within this file.

C. System Definition: Contains modules and connections between these; is abstractly wired with certain interface types.

All three types will be explained below. The example modules show footprints of (1) a Texas Instruments (TI) LaunchPad™ [27] with a 16-bit, ultra-low-power MSP430F5529 MCU [28] and (2) MicroSD card module of type "MicroSD Breakout Board" [29].

```
1  {
2    name: "MSP430F5529_LaunchPad",
3    design: "MSP430F5529_LaunchPad.dbl",
4    interfaces: [{
5        name: "SPI0",
6        type: "SPI",
7        pins: { MISO: "P3.1", MOSI: "P3.0",
8          SCLK: "P3.2", CS: 'any@["P2.0", "P2.2"]' }
9      }, {
10       name: "SPI1",
11       type: "SPI",
12       pins: { MISO: "P4.5", MOSI: "P4.4",
13         SCLK: "P4.0", CS: any }
14     } , {
15       name: "I2C0",
16       type: "I2C",
17       pins: { SDA: "P3.0", SCL: "P3.1" }
18     }, {
19       name: "I2C1",
20       type: "I2C",
21       pins: { SDA: "P4.1", SCL: "P4.2" }
22     }
23   ],
24   pins: ["P6.5", "P3.4", "P3.3", "P1.6",
25   "P6.6", "P3.2", "P2.7", "P4.2", "P4.1",
26   "P6.0", "P6.1", "P6.2", "P6.3", "P6.4",
27   "P7.0", "P3.6", "P3.5", "P2.5", "P2.4",
28   "P1.5", "P1.4", "P1.3", "P1.2", "P4.3",
29   "P4.0", "P3.7", "P8.2", "P2.0", "P2.2",
30   "P7.4", "RST" , "P3.0", "P3.1", "P2.6",
31   "P2.3", "P8.1"]
32 }
```

Figure 4. Module definition of a TI LaunchPad™with two SPI and two I²C interfaces, both overlapping.

A. Module Definitions and Design Blocks

The module definition of (1) a TI LaunchPad™ is shown in Figure 4, whereas the definition of (2) a MicroSD Breakout Board can be seen in Figure 5. Apart from a name and a design block file property, this definition consists of an array of interfaces and pins. The design block file property refers to an EAGLE [30] design block file, comprised of a schematic placeholder (cf. Figures 6 and 7, respectively), and a board layout placeholder (cf. Figures 8 and 9, respectively). These placeholders will later be placed on the output schematics and board layouts. The array of interfaces may contain several different interface types of which the module is capable. The property *type* determines the corresponding interface type. In module (1) in Figure 4, two SPIs and two Inter-Integrated Circuit (I²C, [31]) interfaces are present. Both contain a name, the type (*SPI*, *I2C*), and several pins. Module (2) in Figure 5, on the contrary, is very simple, with only one SPI interface in total.

Pins within interfaces can either be directly assigned to hardware pins (e.g., `MISO: "P3.1"` in line 7, Figure 4) or left for automatic assignment (e.g., `CS: any` in line 13, Figure 4). It is also possible to automatically assign a wire from a dedicated pool by using `any@somearray` (cf. line 8, Figure 4) syntax. This syntax enables the placing of so-called Chip Select (CS) wires in a more detailed way, e.g., based on needs for shorter connection wires, module specifications or other PCB properties. In this case, `somearray` must, of course, be replaced by a JSON-compliant array of strings, being a subset of the pins of the module, cf. Equation (1).

$$somearray \subseteq pins \tag{1}$$

```
1  {
2    name: "MicroSD_BreakoutBoard",
3    design: "MicroSD_BreakoutBoard.dbl",
4    interfaces: [
5      {
6        name: "SPI1",
7        type: "SPI",
8        pins: { MISO: "DO", MOSI: "DI",
9          SCLK: "CLK", CS: "CS" }
10     }
11   ],
12   pins: ["CLK", "DO", "DI", "CS", "CD"]
13 }
```

Figure 5. Module definition of a MicroSD Breakout Board with an SPI interface.

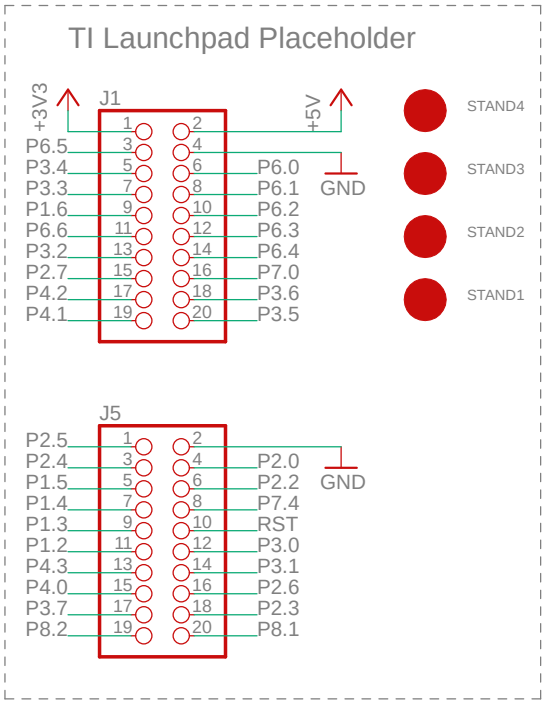


Figure 6. Schematics of a placeholder design block for a TI LaunchPad™ [27].

Each module definition file is associated with its corresponding design block. It is of utmost importance that pin names are coherent in both module representations, as the naming coherence later ensures that proper interconnections are made between modules. Furthermore, a standard format for power supply connections must be used to avoid creating discrepancies between modules. The bus speed of the SPI and the I²C was not considered in this work and will be addressed in future developments towards NFRs. As depicted in Figures 8 and 9, the board layout of a module only consists of its pins. The main idea here was to create a motherboard upon which modules can be placed using their exterior connections (e.g., pin headers or similar connectors). Therefore, the placeholder serves as interface layout between fully assembled PCB modules, such as the LaunchPad™or the Breakout Board, and can then be connected to other modules through interfaces.

B. Interface Definitions

After defining the modules, the generic interfaces must be defined. The interface definition collection is centralized in

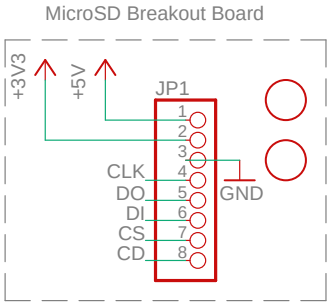


Figure 7. Schematics of a placeholder design block for a MicroSD Breakout Board [29].

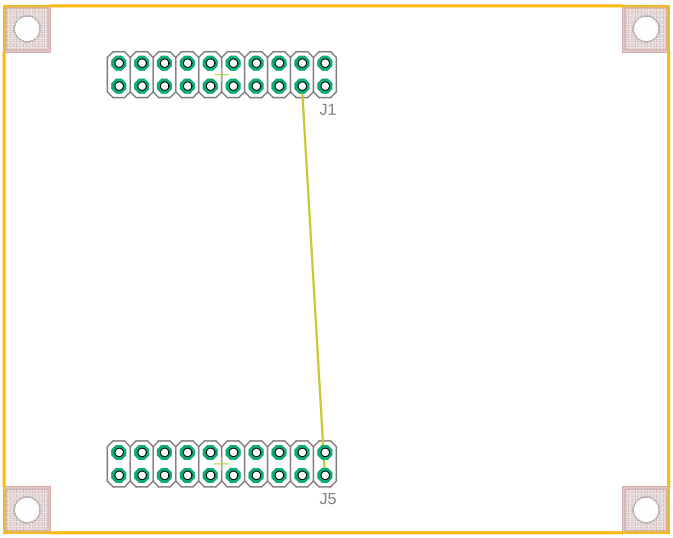


Figure 8. Board layout of a placeholder design block for a TI LaunchPad™ [27].

a single file, and its structure is shown in Figure 10. In this example, only SPI and I²C have been defined with its standard connections. As the format is generic, other interface types, e.g. Controller Area Network (CAN, [32]) or even Advanced eXtensible Interface Bus (AXI, [33]), are also feasible. It also shows how masters and slaves within this communication protocol are connected to the bus wires. As the SPI also has CS wires for every slave selection, special treatment must be used here: A slave only has one CS wire, which is marked with *wiresingle* (cf. line 14, Figure 10), whereas a master has as many CS wires as it has slaves connected to it (marked with *wiremultiple*; cf. line 13, Figure 10). Compared to SPI, the shown example of I²C is rather simple, as it only consists of two wires, with a master/slave concept as well. All participants are simply connected to the corresponding bus wires.

C. System Definition

The final step taken was to define the system itself, which was built from modules and the connections between them. To do so, a single project file must be created, as illustrated in Figure 11. Initially, all necessary modules are imported and named accordingly within the *modules* array. Once defined, they can be interconnected using the previously defined interface definitions. In our example, LaunchPad™ *MSP1* was connected to a MicroSD Breakout Board *SD1* via SPI. This particular SPI connection is called *SPI_Connection1* of type *SPI* and has two participants with different roles: *MSP1* as

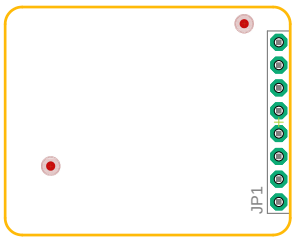


Figure 9. Board layout of a placeholder design block for a MicroSD Breakout Board [29].

```
{
  interfaces: [
    {
      type: "SPI",
      connections: [
        { "master.MOSI" : "bus.MOSI" },
        { "master.MISO" : "bus.MISO" },
        { "master.SCLK" : "bus.SCLK" },
        { "slave.MOSI"  : "bus.MOSI" },
        { "slave.MISO"  : "bus.MISO" },
        { "slave.SCLK"  : "bus.SCLK" },

        { "master.CS"   : "wiremultiple" },
        { "slave.CS"    : "wiresingle" }
      ]
    },
    {
      type: "I2C",
      connections: [
        { "master.SDA" : "bus.SDA" },
        { "master.SCL" : "bus.SCL" },
        { "slave.SDA"  : "bus.SDA" },
        { "slave.SCL"  : "bus.SCL" }
      ]
    }
  ]
}
```

Figure 10. Interface definition containing SPI and I^2C.

a master and *SD1* as a slave. This system definition will prospectively be generated and extracted out of the ASW code by the analysis step in ***papageno***X. The ***papageno***PCB approach is taken to generate PCBs only.

V. Implementation of PCB Generation

After having defined the modules, interfaces, and implemented a system definition, PCB generation can start. The generation consists of two major steps: (*A.*) establishing connection wires based on predefined module and system definitions, and assigning dedicated pins and (*B.*) generating XML-based schematic files from its output. The final step (*C.*), which is carried out to deal with the final layout of the schematics, must be done subsequently (in part manually). The generator is developed as a Java command line application to maintain platform-independence and ensure that it can be integrated into standard tool chains and build management tools.

```
{
  modules: [
    { name: "MSP1",
      type: "MSP430F5529_LaunchPad"},
    { name: "SD1",
      type: "MicroSD_BreakoutBoard"}
  ],
  connections: [
    {
      name: "SPI_Connection1",
      type: "SPI",
      participants: [
        { name: "MSP1", role: "master" },
        { name: "SD1", role: "slave" }
      ]
    }
  ]
}
```

Figure 11. A system model containing two modules connected via SPI.

A. Connection Establishment and Pin Assignment

During this first step, JSON data structure analysis presents the main challenge. The whole system must be interconnected appropriately using the previously explained definition files. To do so, all connections within the system definition must be matched at the beginning of the process. This task subsumes the discovery of connections between modules, their mapping to certain interface types, and the final wire allocation required to interconnect all participants. Specifically, each connection has a type and a finite number of participants with different roles, interfaces, and pins. These pins must then be connected to the newly introduced wires, belonging to the communication. Several different types of wires can be used to connect the participants with each other:

The easiest wires to use are common wires, which can be assigned to a pool of free pins of the module. These wires are marked with *wiresingle* within the interface definition. Due to the fact that all unused General-Purpose Input/Output (GPIO) pins of a module can be used for this purpose, they need to be assigned last.

Furthermore, every participant can connect itself directly to bus wires via its dedicated pins, depending on, e.g., the type of MCU used. In the case of an MSP430 MCU, certain pins are electrically connected to an interface circuit, as defined in its module definition (cf. Figure 4). These pins must, therefore, be matched with the connection's wires (cf. Figure 10). The interface definition must match roles and pins accordingly to correctly interconnect the participants of each connection.

Another type of wires that can be used are multiple wires. If we take SPI as an example, the master needs to have as many chip-select wires as slaves with which it wants to communicate. Therefore, this type of wire – marked with *wiremultiple*, as previously defined – must clone itself to obtain the number of wires needed.

These different types of wires must be connected to the pins of the modules to establish a proper connection or *net* according to the interface definition. The interconnected modules with their nets form a holistic JSON-based description of the system.

B. Schematic and Board Layout Generation

Utilizing the interconnected system description, schematics and board layouts can be generated. In our case, EAGLE's XML data structure [2] was used to form a dedicated output file for schematics and board layouts. To generate those plans, (1) design blocks for each module must be loaded, (2) the previously found connections must be applied and (3) the connected design blocks must be placed on an empty schematic plan or board layout. The basis of every schematic and board plan forms an empty EAGLE plan, on which the explained actions are performed.

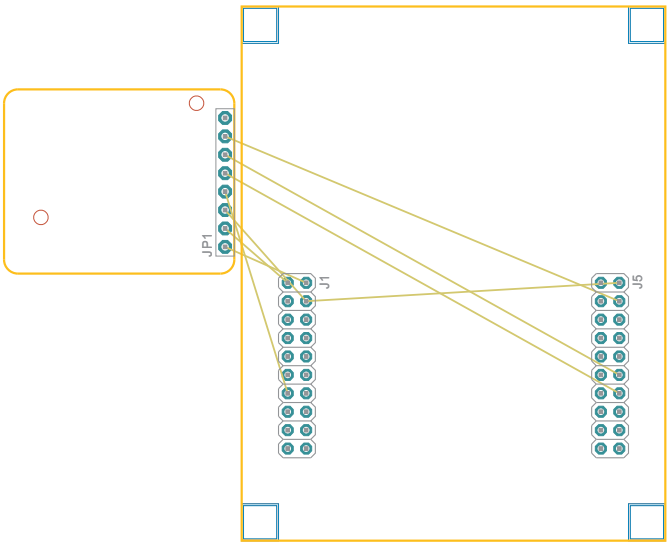


Figure 12. Raw output of the board layout generated as displayed in EAGLE.

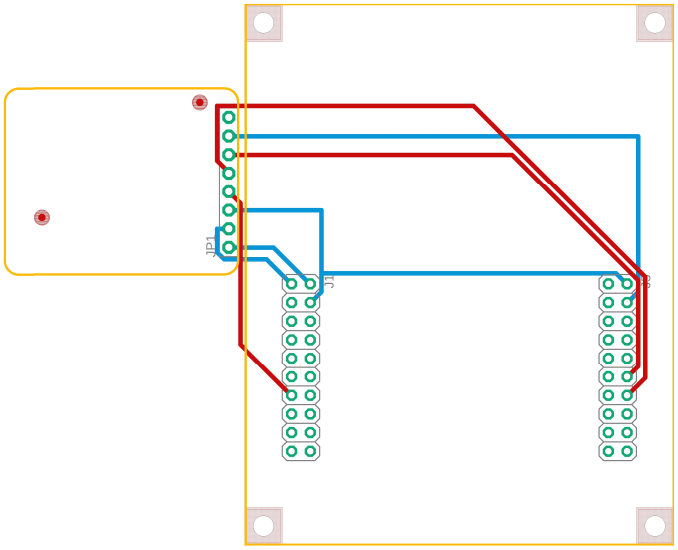


Figure 13. Board layout after auto-routing in EAGLE.

(1) **Instantiation:** In this step, each module must be instantiated by loading the corresponding design block of its type. To avoid overlapping of signal names and, therefore, unwanted connections between modules of the same type, suffixes are added according to the instance's name. For readability purposes, these suffixes are equal to the instance's name defined in the system definition file (e.g., **_MSP1*). This can be easily seen when comparing, e.g., Figures 6/7 and 14.

(2) **Interconnection:** This step must be carried out to form the whole system according to the JSON-based holistic description. Therefore, pins of each module must be assigned to the wires of a connection within the system. To do so, each connection again must be applied separately to each participant. As the system description already contains information, as to which pin of a module must be connected to which wire, this can be done quite easily. In this case, to avoid overlapping of signals, a dot notation style is used do distinguish between wires of different connection instances (e.g., *SPI_CONNECTION1.** in Figure 14).

(3) **Placement:** This step, which is the computationally most expensive step, must be carried out to merge the connected instances of each module into an empty plan, as a great deal of XML parsing is required here. To create consistent plans, the design blocks must be prepared well beforehand to avoid, e.g., inconsistencies within board layers or signal names. To keep the modules from overlapping, a two-dimensional translation of each module must be executed as part of each merge procedure as well. In total, two merging steps are required for each module – one for the schematic and one for the board layout. As this approach generates connection PCBs ("motherboards") where one can plug in modules, only placeholders are used.

Finally, the two generated XML structures are exported and saved into different files (one for the schematics, one for the board layout) for further usage.

C. Routing Generated Schematics and Board Layouts

As layouting and routing of PCBs is a non-trivial task, and engineers need a great deal of experience when performing a task like this, ***papageno***PCB cannot be used to produce final variants of a board. It is recommended to use EAGLE's auto-routing functionality or manual routing to finalize the already well-prepared layouts.

VI. Experiments and Evaluation

Within this section, the previously explained concept on how to define and create PCBs from a definition language is shown in different examples and evaluation. At first, a simple proof of concept is presented in Section VI-A, followed by some analysis and evaluation on scalability and performance of the algorithms in Section VI-B. Section VI-C shows a use case with a corresponding manufactured and equipped prototype PCB. In this case, a comparison with other approaches is not executed, as all related works go in different directions. Hence, there are no acceptable metrics for comparison provided.

A. Proof of Concept

The proof of concept comprises the generation of the system definition as shown in Figure 11. As mentioned above, the system created consists of two modules interconnected with one SPI bus, whereas the processor board serves as master. The schematics generation step yields in the drawing depicted in Figure 14. Compared with the LaunchPad™'s design block shown in Figure 6, one can see the differences in the net names. As examples, *P2.0* has been replaced with *WIRE0*, and *P3.0* is now assigned to *SPI_CONNECTION1.MOSI*. These wires connect to pins 7 and 6 of the MicroSD Breakout Board on the left, respectively. Also, each unconnected pin is given a suffix describing its module (cf. *_MSP1*). These newly introduced net names are the results of the wire generation explained in Section V-A. As the reusability of schematic plans is an important aspect, the feature of non-overlapping module placement can be emphasized as well. The result of the board layout generation step is shown in Figure 12, as described in Section V-B. The fine lines show non-routed connections between the pins. As the generated plan will, of course, be

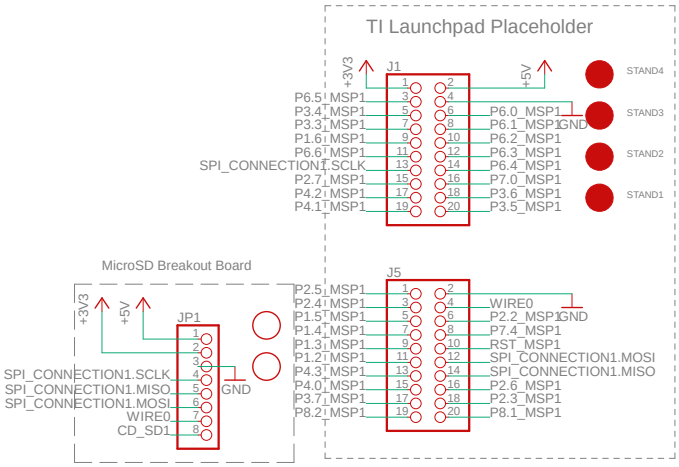


Figure 14. Raw output of the schematics generated as displayed in EAGLE.

manufactured as a real hardware PCB, no single component can overlap in the final layout. Routing of the board has to be either performed manually or by using a design tool's built-in auto router. A feasible layout variant is presented in Figure 13. EAGLE can also be used to check the correctness of the XML file format.

B. Scalability and Performance

In this section, we describe measurements and investigations that concern the performance of the PCB-generating process. All discussed evaluations use one setup as a reference. The application was executed with a Java 10 virtual machine on an Intel Core i7 7500U@2.7GHz with 16 gigabytes of RAM. Table I shows the mean execution time and the combined output XML file size of the generation process of different test case scenarios, which are explained below. All test cases featured a different number of participants (part.) which consisted of masters (M) and slaves (S) with different connection types. The experiment is based on the one presented in [1], but the generation tool version is more stable and optimized and the objective differs slightly, yielding different measurements.

Each test case is based on the example described in Section VI-A, but with different constellations concerning the numbers and types of participants and connections. All test cases were executed 100 times. Four types of test scenarios with seven test cases each were conducted: Within the first scenario, just one SPI connection was present, with a varying number of slaves each test case. The second scenario comprised two SPI connections with an increasing number of slaves. Test scenario three had one I^2C connection and was similar to scenario one, whereas scenario four included SPI and I^2C connections to a single master with an increasing number of slaves. The devolution of the mean execution time (in ms) in all test scenarios is shown in Figure 15. When comparing all scenarios, the trend observed is relatively similar: All performance graphs show a linear devolution with an additive, logarithmic-like component. The linear component is due to the linear increase in the complexity of the test cases. The logarithmic-like growth observed can be explained by the decreasing, additive overhead of the linear component when processing similar connection reasoning, as well as the XML schematic and layout data. This is also the reason why

TABLE I. MEAN EXECUTION TIMES FOR DIFFERENT SCENARIOS.

#	*test scenario description*	*execution time*	*file size*
	1 SPI conn. (scenario 1)		
0	2 part. (1 M, 1 S)	656.22 ms	94 KiB
1	3 part. (1 M, 2 S)	751.98 ms	111 KiB
2	4 part. (1 M, 3 S)	852.95 ms	128 KiB
3	5 part. (1 M, 4 S)	933.71 ms	144 KiB
4	6 part. (1 M, 5 S)	1 013.67 ms	161 KiB
5	7 part. (1 M, 6 S)	1 108.81 ms	178 KiB
6	7 part. (1 M, 7 S)	1 193.25 ms	194 KiB
	2 SPI conn. (scenario 2)		
0	4 part. (1 M and 1 S each)	890.77 ms	160 KiB
1	6 part. (1 M and 2 S each)	1 060.91 ms	192 KiB
2	8 part. (1 M and 3 S each)	1 234.36 ms	226 KiB
3	10 part. (1 M and 4 S each)	1 398.30 ms	259 KiB
4	12 part. (1 M and 5 S each)	1 557.30 ms	293 KiB
5	14 part. (1 M and 6 S each)	1 731.26 ms	326 KiB
6	14 part. (1 M and 7 S each)	1 879.93 ms	360 KiB
	1 I^2C conn. (scenario 3)		
0	2 part. (1 M, 1 S)	665.61 ms	106 KiB
1	3 part. (1 M, 2 S)	771.77 ms	136 KiB
2	4 part. (1 M, 3 S)	889.56 ms	167 KiB
3	5 part. (1 M, 4 S)	989.33 ms	198 KiB
4	6 part. (1 M, 5 S)	1 106.84 ms	228 KiB
5	7 part. (1 M, 6 S)	1 205.48 ms	259 KiB
6	7 part. (1 M, 7 S)	1 332.82 ms	290 KiB
	1 I^2C and 1 SPI conn. (scenario 4)		
0	3 part. (1 M, 1 S each)	782.70 ms	127 KiB
1	5 part. (1 M, 2 S each)	971.42 ms	174 KiB
2	7 part. (1 M, 3 S each)	1 166.01 ms	222 KiB
3	9 part. (1 M, 4 S each)	1 344.08 ms	269 KiB
4	11 part. (1 M, 5 S each)	1 530.92 ms	317 KiB
5	13 part. (1 M, 6 S each)	1 719.09 ms	364 KiB
6	13 part. (1 M, 7 S each)	1 873.92 ms	411 KiB

doubling the numbers in the first test case resulted in much higher values than in test case two. Test scenario four is the only one that displays a steeper curve. This is due to the combination of different connection types, yielding less-optimal algorithm executions. As XML processing is quite costly, some further optimizations are needed. As the overall file size displayed linear growth, no correlation was observed between file size and execution time.

C. Simple Use Case and Prototype Manufacturing

The simple use case in this section is a minimalistic, generic control system. In order to react to its environment, it must

- read several analog voltage values,
- output analog voltage values, and
- store a data log permanently in two different ways, such as it is on the one hand side
  - "detachable", and on the other hand side
  - stored with low energy consumption, yet non-volatile and redundant.

After manually spanning a selection space over the available equipment in our lab, those requirements yield in a system configuration with

- an MCU to execute the control algorithms ($\rightarrow$ MSP430 on a corresponding LaunchPad™),
- two 4-channel analog-to-digital converters (ADCs) to read voltage values ($\rightarrow$ Adafruit 4-channel Breakout Board featuring an ADS1115 ADC [34]),
- a digital-to-analog converter (DAC) to output voltage values ($\rightarrow$ Adafruit 12-bit DAC board featuring a MCP4725 DAC [35]),

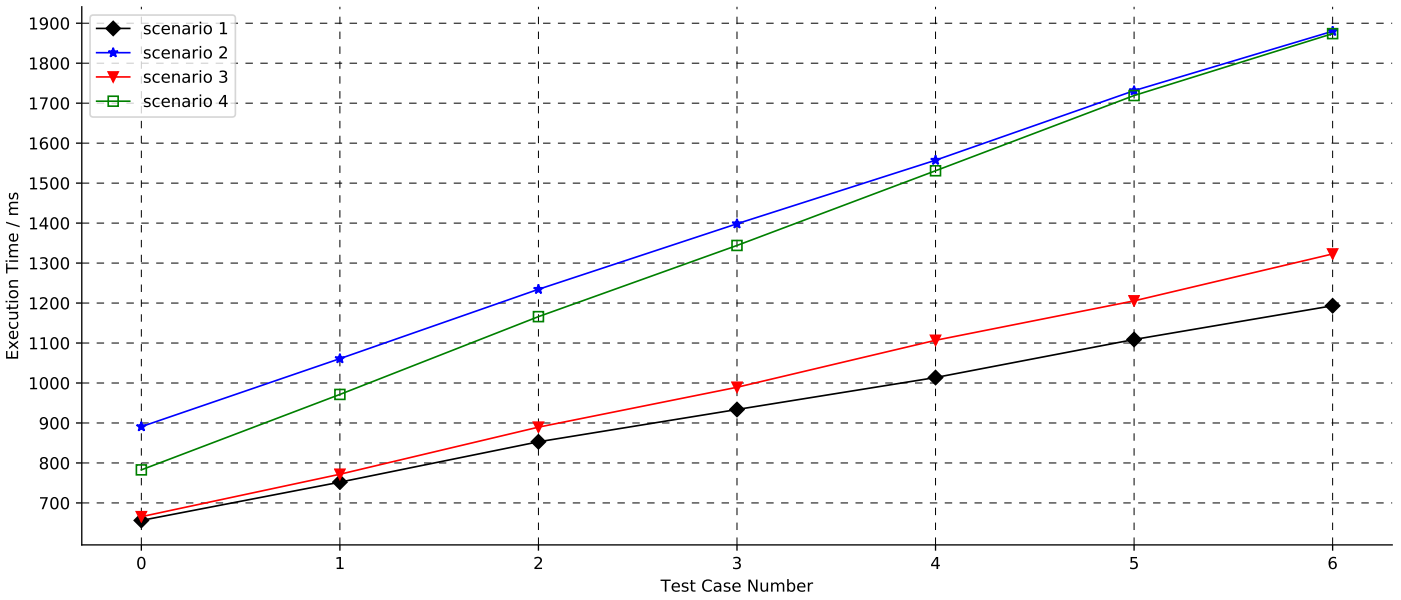


Figure 15. Performance graph for different test cases in all four scenarios.

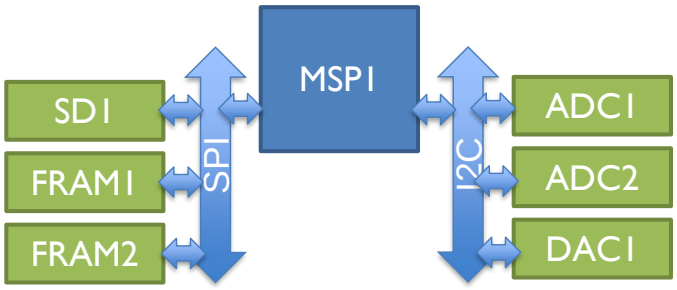


Figure 16. The block diagram of all modules in the example use case.

- a MicroSD card module to log data in a detachable way (→ MicroSD Breakout Board), and
- two Ferroelectric Random Access Memory (FRAM, [36]) modules to store data in a low-power, non-volatile, redundant way (→ Adafruit SPI FRAM Breakout Board featuring a MB85RS64V FRAM [37]).

The block diagram of this configuration, including its interconnection, is shown in Figure 16. It can be seen that a total of two different connection types must be used to interconnect all modules. The corresponding system definition is presented in Figure 17. It contains all module instances, both connections with their types, participants, and roles. The result of running the PCB generation and manually routing the board layout is depicted in Figure 18. With this result it is possible to manufacture an actual PCB, equip it with the hardware modules, flash the control system ASW and BSW, and run measurements. The software setup in this case consists of our own real-time operating system *MCSmart*OS [38][39] enriched with a modular driver management system and a simple test application. The equipped and running prototype is shown in Figure 19, where it is connected to several measurement devices (e.g., a PicoScope 2205 MSO [40] with digital and analog inputs) through debug wires and probes to observe and verify correct functionality.

```
{
  modules: [
    { name: "MSP1",  type: "MSP430F5529_Launchpad" },
    { name: "ADC1",
      type: "Adafruit_ADS1115_16Bit_I2C_ADC" },
    { name: "ADC2",
      type: "Adafruit_ADS1115_16Bit_I2C_ADC" },
    { name: "DAC1",
      type: "Adafruit_MCP4725_12Bit_I2C_DAC" },
    { name: "FRAM1", type: "Adafruit_FRAM_SPI" },
    { name: "FRAM2", type: "Adafruit_FRAM_SPI" },
    { name: "SD1",   type: "MicroSD_BreakoutBoard" }
  ],

  connections: [
    {
      name: "I2C_Connection1",
      type: "I2C",
      participants: [
        { name: "MSP1", role: "master" },
        { name: "ADC1", role: "slave" },
        { name: "ADC2", role: "slave" },
        { name: "DAC1", role: "slave" }
      ]
    },
    {
      name: "SPI_Connection1",
      type: "SPI",
      participants: [
        { name: "MSP1",  role: "master" },
        { name: "FRAM1", role: "slave" },
        { name: "FRAM2", role: "slave" },
        { name: "SD1",   role: "slave" }
      ]
    }
  ]
}
```

Figure 17. The system model of the prototype.

VII. Conclusion and Future Work

In conclusion, the present work based on the ***papageno***PCB approach represents a novel, top-down concept to develop an embedded system for a multitude of possible application scopes. Having only a model-based system description at hand, it is possible to use ***papageno***PCB to generate hardware schematics and board layouts accordingly. This opens up numerous new possibilities towards automatic system generation

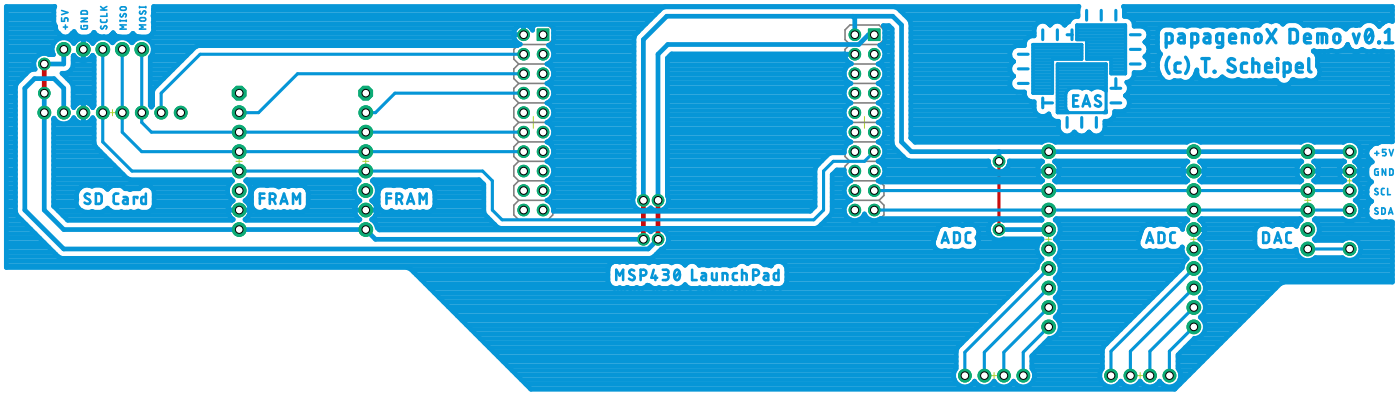


Figure 18. Prototype PCB layout with a MSP430 LaunchPad™ connected to three SPI and three I²C modules (manually routed).

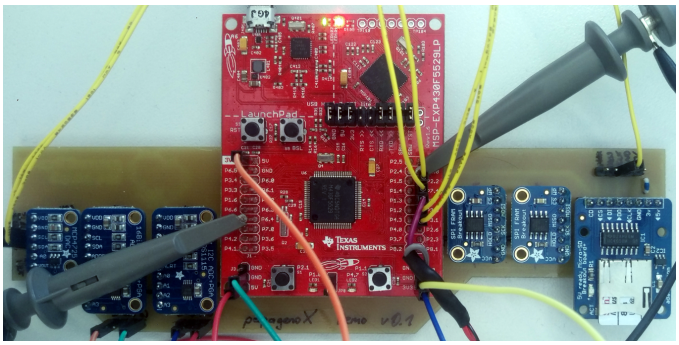


Figure 19. Fully equipped prototype board with debug wires.

and across several abstraction levels including, e.g., automatic bus balancing, bandwidth engineering, optimization towards functional and non-functional hardware requirements. All these things can be carried out even before building the actual hardware for the system. The use of these concepts requires the availability of in-depth information about the electrical and mechanical characteristics of all parts of a PCB, so that the hardware can be optimized in terms of non-functional metrics such as bandwidth or power consumption. Generally speaking, the presented concept is able to optimize systems under development regarding different, user-defined metrics already at design level. Therefore, metrics to measure the overall improvement in general are hard to define, as they depend on the actual system's development process and its requirements and properties. Due to the generic design, new models can be integrated easily, and it will be even possible to take a non-module-based approach on the electrical device or component level, proper definitions presumed.

Concerning future work, a detailed extraction of system models from a profound ASW source code analysis is of utmost importance. Therefore, we are working on introducing annotations into our operating system environment [39], which will enable us to automatically generate system definition files. These annotations can either be introduced into the code as compiler keywords (e.g., pragmas, defines) or as comments. As some work is already being done to improve the automatic portability of real-time operating systems [41], the proposed approach could be used to build a system for which only the application code must be programmed. The rest of the system can then be generated automatically. Even suitable and application-optimized processor architectures [42] or application-specific logic components on reconfigurable computing platforms could be created and included by taking this approach. The ultimate goal is to establish ***papageno***X as a universal embedded systems generator, which uses only ASW source code or models as an input.

References

[1] T. Scheipel and M. Baunach, "papagenoPCB: An Automated Printed Circuit Board Generation Approach for Embedded Systems Prototyping," in *ICONS 2019 - The Fourteenth International Conference on Systems*, 3 2019, pp. 20–25.

[2] Autodesk, Inc., *EAGLE XML Data Structure 9.1.0*, 2018.

[3] J. Schäuffele and T. Zurawka, *Automotive Software Engineering*, ser. ATZ/MTZ-Fachbuch. Springer Fachmedien Wiesbaden, 2016.

[4] A. Kouba, J. Navratil, and B. Hnilička, "Engine Control using a Real-Time 1D Engine Model," in *VPC – Simulation und Test 2015*, J. Liebl and C. Beidl, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2018, pp. 295–309.

[5] Infineon Technologies AG, "TC1797 – 32-Bit Single-Chip Microcontroller," 2014.

[6] B. Eichberger, E. Unger, and M. Oswald, "Design of a versatile rapid prototyping engine management system," in *Proceedings of the FISITA 2012 World Automotive Congress*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 135–142.

[7] Infineon Technologies AG, "TC1796 – 32-Bit Single-Chip Microcontroller," 2007.

[8] ——, "TC1798 – 32-Bit Single-Chip Microcontroller," 2014.

[9] L. H. Crockett, R. A. Elliot, M. A. Enderwitz, and R. W. Stewart, *The Zynq Book: Embedded Processing with the Arm Cortex-A9 on the Xilinx Zynq-7000 All Programmable Soc*. UK: Strathclyde Academic Media, 2014.

[10] devicetree.org, *Devicetree Specification*, Dec. 2017, release v0.2.

[11] G. M. Swinkels and L. Hafer, "Schematic generation with an expert system," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 9, no. 12, pp. 1289–1306, Dec 1990.

[12] B. Singh, D. O'Riordan, B. G. Arsintescu, A. Goel, and D. R. Deshpande, "System and method for circuit schematic generation," US Patent US7 917 877B2, 2011.

[13] J. Schnerr, O. Bringmann, A. Viehl, and W. Rosenstiel, "High-performance Timing Simulation of Embedded Software," in *2008 45th ACM/IEEE Design Automation Conference*, June 2008, pp. 290–295.

[14] B. Schommer, C. Cullmann, G. Gebhard, X. Leroy, M. Schmidt, and S. Wegener, "Embedded Program Annotations for WCET Analysis," in *WCET 2018: 18th International Workshop on Worst-Case Execution Time Analysis*. Barcelona, Spain: Dagstuhl Publishing, Jul. 2018, [retrieved: Nov, 2019]. [Online]. Available: https://hal.inria.fr/hal-01848686

[15] S. Chakravarty, Z. Zhao, and A. Gerstlauer, "Automated, retargetable back-annotation for host compiled performance and power modeling," in *9th Int'l Conference on Hardware/Software Codesign and System Synthesis*, Piscataway, NJ, USA, 2013, pp. 36:1–36:10, [retrieved: Nov, 2019]. [Online]. Available: http://dl.acm.org/citation.cfm?id=2555692.2555728

[16] A. D. Pimentel, "Exploring Exploration: A Tutorial Introduction to Embedded Systems Design Space Exploration," *IEEE Design Test*, vol. 34, no. 1, pp. 77–90, Feb 2017.

[17] F. Herrera, H. Posadas, P. Peñil, E. Villar, F. Ferrero, R. Valencia, and G. Palermo, "The COMPLEX methodology for UML/MARTE Modeling and design space exploration of embedded systems," *Journal of Systems Architecture*, vol. 60, no. 1, pp. 55 – 78, 2014, [retrieved: Nov, 2019]. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S138376211300194X

[18] T. Saxena and G. Karsai, "A meta-framework for design space exploration," in *2011 18th IEEE International Conference and Workshops on Engineering of Computer-Based Systems*, April 2011, pp. 71–80.

[19] N. Nethercote, P. J. Stuckey, R. Becket, S. Brand, G. J. Duck, and G. Tack, "MiniZinc: Towards a Standard CP Modelling Language," in *Principles and Practice of Constraint Programming – CP 2007*, C. Bessière, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 529–543.

[20] AUTOSAR, "Classic platform release 4.3.1," 2017.

[21] IEEE Standards Association, *IEEE 1076-2008 - IEEE Standard VHDL Language Reference Manual*, 2008.

[22] ——, *IEEE 1364-2005 - IEEE Standard for Verilog Hardware Description Language*, 2005.

[23] ——, *IEEE 1666-2011 - IEEE Standard for Standard SystemC Language*, Sep. 2011.

[24] B. Maes, "Comparison of contemporary file systems," *Citeseer*, 2012.

[25] S. C. Hill, J. Jelemensky, M. R. Heene, S. E. Groves, and D. N. Debrito, "Queued serial peripheral interface for use in a data processing system," US Patent US4 816 996, 1989.

[26] ECMA International, *ECMA-404: The JSON Data Interchange Syntax*, 2nd ed., Dec. 2017.

[27] Texas Instruments, *MSP430F5529 LaunchPad™ Development Kit (MSP--EXP430F5529LP)*, Apr. 2017.

[28] ——, *MSP430x5xx and MSP430x6xx Family User's Guide*, Mar. 2018, [retrieved: Nov, 2019]. [Online]. Available: http://www.ti.com/lit/ug/slau208q/slau208q.pdf

[29] Adafruit Industries, *Micro SD Card Breakout Board Tutorial*, Jan. 2019, [retrieved: Nov, 2019]. [Online]. Available: https://cdn-learn.adafruit.com/downloads/pdf/adafruit-micro-sd-breakout-board-card-tutorial.pdf

[30] Autodesk, Inc., "EAGLE," [retrieved: Nov, 2019]. [Online]. Available: https://www.autodesk.com/products/eagle/

[31] NXP Semiconductors, Inc., *UM10204: I2C-bus specification and user manual*, Apr. 2014, rev. 6.

[32] International Organization for Standardization, *ISO 11898: Road vehicles – Controller area network (CAN)* , 2nd ed., Dec. 2015.

[33] ARM Ltd., *AMBA AXI and ACE Protocol Specification*, 2017, [retrieved: Jul, 2019].

[34] Texas Instruments, *Ultra-Small, Low-Power,16-Bit Analog-to-Digital Converter with Internal Reference*, Oct. 2009, [retrieved: Nov, 2019]. [Online]. Available: http://www.ti.com/lit/ds/symlink/ads1114.pdf

[35] Microchip, *12-Bit Digital-to-Analog Converter with EEPROM Memory in SOT-23-6*, 2009, [retrieved: Nov, 2019]. [Online]. Available: https://cdn-shop.adafruit.com/datasheets/mcp4725.pdf

[36] H. Ishiwara, M. Okuyama, and Y. Arimoto, *Ferroelectric random access memories: fundamentals and applications*. Springer Science & Business Media, 2004, vol. 93.

[37] Fujitsu Semiconductor, *64KBit SPIMB85RS64V*, 2013, [retrieved: Nov, 2019]. [Online]. Available: https://cdn-shop.adafruit.com/datasheets/MB85RS64V-DS501-00015-4v0-E.pdf

[38] M. Baunach, "Advances in Distributed Real-Time Sensor/Actuator Systems Operation," Dissertation, University of Würzburg, Germany, Feb. 2013. [Online]. Available: http://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/6429

[39] R. Martins Gomes, M. Baunach, M. Malenko, L. Batista Ribeiro, and F. Mauroner, "A Co-Designed RTOS and MCU Concept for Dynamically Composed Embedded Systems," in *Proc. of the 13th Workshop on Operating Systems Platforms for Embedded Real-Time Applications*, 2017, pp. 41–46.

[40] *PicoScope 2205 MSO Mixed Signal Oscilloscope*, Pico Technology, 2016. [Online]. Available: https://www.picotech.com/download/datasheets/PicoScope2205MSODatasheet-en.pdf

[41] R. Martins Gomes and M. Baunach, "A Model-Based Concept for RTOS Portability," in *Proc. of the 15th Int'l Conference on Computer Systems and Applications*, Oct. 2018, pp. 1–6.

[42] F. Mauroner and M. Baunach, "mosartMCU: Multi-Core Operating-System-Aware Real-Time Microcontroller," in *Proc. of the 7th Mediterranean Conference on Embedded Computing*, Jun. 2018, pp. 1–4.

Anomaly Detection and Analysis for Reliability Management in Clustered Container Architectures

Areeg Samir, Nabil El Ioini, Ilenia Fronza, Hamid R. Barzegar, Van Thanh Le and Claus Pahl

Faculty of Computer Science
Free University of Bozen-Bolzano
39100 Bolzano, Italy
Email: firstname.surname@unibz.it

***Abstract*—Virtualised environments such as cloud and edge computing architectures allow software to be deployed and managed through third-party provided services. Here virtualised resources available can be adjusted, even dynamically to changing needs. However, the problem is often the boundary between the service provider and the service consumer. Often there is no direct access to execution parameters at resource level on the provider's side. Generally, only some quality factors can be directly observed while others remain hidden from the consumer. We propose an architecture for autonomous anomaly analysis for clustered cloud or edge resources. The key contribution is that the architecture determines possible causes of consumer-observed anomalies in an underlying provider-controlled infrastructure. We use Hidden Hierarchical Markov Models to map observed performance anomalies to hidden resources, and to identify the root causes of the observed anomalies in order to improve reliability. We apply the model to clustered hierarchically organised cloud computing resources. We illustrate use cases in the context of container technologies to show the utility of the proposed architecture.**

***Index Terms*—Cloud Computing; Edge Computing; Container Technology; Cluster Architectures; Markov Model; Anomaly Detection; Performance; Workload.**

I. Introduction

As a consequence of the dynamic nature of cloud and edge computing environments, users may experience anomalies in performance caused by the distributed nature of clusters, heterogeneity, or scale of computation on underlying resources that may lead to performance degradation and application failure, for example

- change in a cluster node workload demand or configuration updates may cause dynamic changes,
- reallocation or removal of resources may affect the workload of system components.

In principle, application deployments can be adjusted, even dynamically to changing conditions. A problem, however, emerges. Cloud and edge computing allow applications to be deployed in remote environments, but these are managed by third parties based on provided virtualised resources [1], [2], [3], [4] which often hides the underlying causes from the consumers of these services.

In virtualised environments, some factors can be directly observed (e.g., application performance) while others remain hidden from the consumer (e.g., reason behind the workload changes, the possibility of predicting the future load, dependencies between affected nodes and their load). Thus, the reasons for these anomalies remain unclear. Recent works on anomaly detection [5], [6], [7] have looked at resource usage, rejuvenation or analysing the correlation between resource consumption and abnormal behaviour of applications. However, more work is needed on identifying the reason behind observed resource performance degradations.

We here investigate the possible root causes of performance anomalies in an underlying provider-controlled cloud infrastructure. We propose an anomaly detection and analysis architecture for clustered cloud and edge environments that aims at automatically detecting possibly workload-induced performance fluctuations, thus improving the reliability of these architectures. System workload states that might be hidden from the consumer may represent anomalous or faulty behaviour that occurs at a point in time or lasts for a period of time. An anomaly may represent undesired behaviour such as overload, or also appreciated positive behaviour like underload (the latter can be used to reduce the load from overloaded resources in the cluster). Emissions from those states (i.e., observations) indicate the possible occurrence of failure resulting from a hidden anomalous state (e.g., high response time). In order to link observations and the hidden states, we use Hierarchical Hidden Markov Models (HHMMs) [10] to map the observed failure behaviour of a system resource to its hidden anomaly causes (e.g., overload) in a hierarchically organised clustered resource configuration. Hierarchies emerge as a consequence of a layered cluster architecture that we assume based on a clustered cloud computing environment. We aim to investigate, how to analyse anomalous resource behaviour in clusters consisting of nodes with application containers as their load from a sequence of observations emitted by the resource.

We focus on a clustered, hierarchically organised environment with containers as loads on the individual nodes, similar to container cluster solutions like Docker Swarm or Kubernetes [36]. We use a detailed use case discussion in container technologies to illustrate the applicability of the proposed solution.

In order to broaden the discussion, we also expand our anomaly notion. In addition to performance and workload

anomalies, we introduce trust anomalies and discuss the transferability of concepts to this trust concern.

This paper is structured as follows. Section II discusses the state of the art. Section III introduces our wider anomaly management architecture. Section IV details the anomaly detection and fault analysis. Section V evaluates the proposed architecture. This is followed by an extended use case discussion in Section VI that shows the applicability of the results. Section VII discusses the transferability of concerns to a trust anomaly context. Section VIII ends the paper with some conclusions and possible future work.

II. Related Work

This section explores the detection, identification, and recovery of anomaly in literature. Moreover, it sheds light on the literatures that use the Hidden Markov Model to mitigate the anomalous behavior.

A. Anomaly Detection and Identification

Several studies [11] and [7] have addressed workload analysis in dynamic environments. Sorkunlu et al. [12] identify system performance anomalies through analysing the correlations in the resource usage data. Peiris et al. [13] analyse the root causes of performance anomalies by combining the correlation and comparative analysis techniques in distributed environments.

Dullmann et al. [14] provide an online performance anomaly detection approach that detects anomalies in performance data based on a discrete time series analysis. Wang et al. [7] model the correlation between workload and the resource utilization of applications to characterize the system status. However, the authors work neither classifies the different types of workloads, or recovers the anomalous behaviour.

In [26] the author detects the anomalous behaviours (CPU overload and Denial of Service Attack), and provides an adaptation policy using a multi-dimensional utility-based model and algorithms. The author gives a score and likelihood for the anomaly detected to select an adaptation policy to be able to scale compute resources. The author work specifies a node leader for each microservice cluster. Each node maintains the cluster state and preserves the cluster logs. The leader also votes on the adaptation policy action. However, the author work handles two types of anomalies, and it is limited to the horizontal and vertical auto-scaling actions to mitigate the anomalous behaviour. Further, the work does not predict the future workload.

The work in [46] detects the anomalous behaviour in performance using a forecasting model to estimate the bandwidth, detect performance changes and to decompose time series into components. However, the authors use a hard threshold in all the dataset which may not reflect the actual workloads in system. In addition, they only detect anomalies without analysing them, and they use labelled-time which is not good enough to detect all anomalies as some anomalies could not be discovered during the detection process and time complexity in terms of data size may occur.

In [38] the authors focus on detecting anomalous behaviour of services deployed on VM in a cloud environment. Like our architecture, different anomaly injection scenarios are created and a workload is generated to test the impact of anomaly on the cloud services. The authors emulated different anomalies with the CPU, memory, disk, and network. However, their work does not track the cause of anomalous behaviour in a containerized cluster environment, and it neglects the dependency between nodes.

The work in [50] implements a probabilistic prediction model based on a supervised learning method. The model aims at detecting anomalous behaviour in cloud infrastructure through analysing correlation between different metrics (CPU, memory, disk, and network) to find the essential metrics that can characterize the correlation between cloud performance and anomaly event. The work uses a directed acyclic graph to analyse the correlation of various performance metrics with failure events in a virtual and physical machines. The author computes the conditional probability of every metric on anomaly occurrences, and selects those metrics whose conditional probabilities are greater than a predefined threshold. Nevertheless, the results show that the model suffers from poor prediction efficiency when it is used to predict cloud anomalies.

The work in [54] presents a general purpose prediction model to prevent anomalies in cloud environment. The author uses a supervised learning-based model that combines two dependent Markov chain models with the tree augmented Bayesian networks. The work applies statistical learning algorithms over system level metrics (CPU, memory, network I/O statistics) to predict the anomalous behaviour. However, the author does not provide information about the prediction efficiency.

The work in [61] predicts the impact of processor cache interference among consolidated workloads at application level. To predict the performance degradation of consolidated applications, the prediction technique is only linear to the number of cores sharing the last-level cache. However, the author limits its discussion to cache contention issues, ignoring other resource types.

The work in [47] develops a description language ”Performance Problem Diagnostics Description Model” to specify the information required for conducting an automatic performance problem diagnostics. The work analyses the workload to detect and categorize the faults into three layers namely. (1) Symptoms, externally visible indicators of a performance problem, (2) manifestation, internal performance indicators or evidences, and (3) root-causes, physical factors whose removal eliminates the manifestations and symptoms of a performance incident. However, the approach neither considers the dependency between faults nor avoids human interaction (i.e., performance experts should provide heuristics to be able to detect performance problems). The approach is designed to

apply for a specific domain, it does not provide a recovery mechanism to the detected faults neither discovers the dependency between anomalies. Further the approach is based on predefined heuristics (rules) to detect performance problems. Consequently, applying the approach on a different domain or changing the fault model requires heuristics update.

The work in [70] proposes an approach for localizing anomalies at operation time of a target system using the Kieker monitoring approach. For the localization of anomalies, the author calculates an anomaly score for an operation through specifying a threshold. The author specifies a set of rules to detect performance anomaly. The rules are continuously evaluated based on the anomaly score through using forecasting techniques to predict future values in a time series. The author evaluates the observed measurement values, (i.e. response times) with the forecasted values to detect anomalies. However, the work ignores the type of the performance anomaly and anomaly dependency.

The work in [39] localizes faulty resources in cloud environments through modelling correlations among anomalous resources. The author uses the graph theory to locate the correlation between pairs of resources. The author focuses on analysing the amount of occupied memory in a physical server, the CPU consumption of a virtual machine, and the number of connections accepted by an application. However, the author work does not target anomaly in microservice or container.

In [30] the author studies the performance of several machine learning models to predict attacks on the IoT systems accurately. The results show that the random forest model achieves a promising anomaly prediction comparing to the other machine learning models. Nevertheless, the work only concentrates on predicting the network anomaly.

In [34] the author proposes an approach to estimate the capacity of a microservice by measuring the maximal number of successfully processed user requests per second for a given service such that no Service Level Objective SLO is violated. The author conducts a limited set of load tests followed by fitting an appropriate regression model to the acquired performance data. The author work examines the impact of workload on the CPU and memory usage. The author mentions that changing the number of requests affects the number of virtual CPU cores but it does not affect the memory utilization significantly. However, the work does not predict the future workload. Also, the work neglects the dependency between the nodes and services.

The work in [42] investigates the network performance impact of containers deployed on virtual machines. The author does several experiments to analyse the network performance of containers considering the horizontal scaling and network data transfer rate. Nevertheless, the work concentrates only on network and its impact on container performance.

In [43] the work explores the affect of microservices on each other on the same host. The author measures the CPU, memory and network usage metrics of the containers and nodes. However, the work is limited to evaluate the current failure prediction methods in Microservice environment. Moreover, the work does not locate or detect anomalous behaviour, and it focuses is CPU-bound workload.

B. Hidden Markov Model

Many literatures use the HMM, and its derivations to detect anomaly. In [17], the author proposes various techniques implemented for the detection of anomalies and intrusions in the network using the HMM.

Ge et al. [19] detect faults in real-time embedded systems. The authors use the HMM to describe the healthy and faulty states of a systems hardware components. In [22] the HMM is used to find which anomaly is part of the same anomaly injection scenarios.

C. Anomaly Recovery

In [28] the author provides a fault tolerance management mechanism at the Physical Machines and Virtual Machines levels. the work uses Redundant Array of Independent Disks (RAID-6) to optimize the space storage and to recover data in case of machine failure. The author divides a set of VM and PM into sub-sets of the same size. The author uses two services to gather information about a resource status and to manage resources through adding and deleting resources to mitigate resource failure. Nevertheless, the author only focuses on two aspects of recovery: handle the storage disk crash, and deal with the operating system crash.

Maurya and Ahmad [16] propose an algorithm that dynamically estimates the load of each node and migrates the task on the basis of predefined constraint [31]. However, the algorithm migrates the jobs from the overloaded nodes to the underloaded one through working on pair of nodes, it uses a server node as a hub to transfer the load information in the network which may result in overhead at the node.

In [77] the author presents a control theory-based consolidation approach that mitigates the effects of the cache, memory and hardware contention of coexisting workloads. The approach manages interference among consolidated VMs by dynamically allocating the resources to applications based on the workload SLAs. But, the author focuses is CPU-bound workload and compute-intensive applications.

III. Self-Adaptive Fault Management

Our ultimate goal is a self-adaptive fault management architecture [9], [8], [23] for cloud and edge computing that automatically identifies anomalies by locating the reasons for degradations of performance, and making explicit the dependency between observed failures and possible faults cause by the underlying cloud resources.

A. The Fault Management Framework

Our complete architecture consists of two models: (1) Fault management model that detects and identifies anomalies within the cloud system. (2) Recovery model that applies a recovery mechanism considering the type of the detected anomaly and the resource capacity. Figure 1 presents the overall architecture. The focus in this paper is on the Fault management model.

The cloud resources consist of a cluster, which composed of a set of nodes that host application containers as loads deployed on them. Each node has an agent that can deploy containers and discover container properties. We use the container notion to embody some basic principles of container cluster solutions [15] such as the Docker Swarm or Kubernetes, to which we aim to apply our architecture ultimately.

We align the architecture with the Monitor, Analysis, Plan, Execute based on the anomaly detection Knowledge (MAPE-K) feedback control loop. The Monitor, collects data regarding the performance of the system as the observable state of each resource [18]. This can later be used to compare the detected status with the currently observed one. Each anomalous state has a weight (probability of occurrence). The identification step is followed by the detection to locate the root cause of anomaly (Analysis and Plan). The identified anomalous state is added to a queue that is ordered based on its assigned weight to signify urgency of healing. The Knowledge about anomalous states are kept on record. Different recovery strategies (Execute) can mitigate the detected anomalies. Different pre-defined thresholds for recovery activities are assigned to each anomaly category based on the observed response time failures. Corresponding rules can be updated with the results from the recovery stage. This update aids in learning our models and enhancing the future detection.

The detection of an anomaly is based on using historical performance data to determine probabilities. We classify system data into two categories. The first one reflects observed system failures (essentially regarding permitted response time), and the second one indicates the (hidden) system faults related to workload fluctuations (e.g., by containers consuming a resource). We further annotate each behavioural category to reflect the severity of anomalous behaviour within the system, and the probability of its occurrence. The response time behaviour captures the amount of time taken from sending a request until receiving the response (e.g., creating container(s) within a node). For example, observed response time can fluctuate. The classified response time should be linked to the failure behaviour within the system resources (i.e., CPU) to address unreliable behaviour. We can also classify the resource workload into normal load (NL), overload (OL), and underload (UL) categories to capture the workload fluctuations.

B. Anomaly Detection and Identification

Anomaly detection, the Monitoring stage in the MAPE-K, collects and classifies system data. It compares new collected data with previous observations based on the specified rules in the Knowledge component.

Fault identification, the Analysis and Plan stages in the MAPE-K, identifies the fault type and its root cause to explain the anomalous behaviour. The main aim of this step is specifying the dependency between faults (the proliferation of an anomaly within the managed resources), e.g., an inactive container can cause another container to wait for input. We use the Hierarchical Hidden Markov models (HHMM) [10], a doubly stochastic model for modeling hierarchical structures of data, to identify the source of anomalies.

Based on the response time emissions, we trace the path of the observed states in each observation window. Once we diagnose anomalous behaviour, the affected nodes are annotated with a weight, which is a probability of fault occurrence for an observed performance anomaly. Nodes are addressed based on a first-detected-first-healed basis.

In order to illustrate the usefulness of this analysis, we also discuss the fault handling and recovery in the next subsection. Afterwards, we define the HHMM model structure and the analysis process in detail.

C. Fault Handling and Recovery

After detecting and identifying faults, a recovery mechanism, the Execute stage in the MAPE-K, is applied to carry out the load balancing or the other suitable remedial actions, aiming to improve resource utilization. Based on the type of the fault, we apply a recovery mechanism that considers the dependency between nodes and containers. The recovery mechanism is based on current and historic observations of response time for a container as well as knowledge about the hidden states (containers or nodes) that might have been learned.

The objective of this step is to self-heal the affected resource. The recovery step receives an ordered weighted list of faulty states. The assigned probability of each state based on a predefined threshold is used to identify the right healing mechanism, e.g., to achieve fair workload distribution. Once a state has recovered, it is removed from an anomaly queue, stored it in the recovered list flagged as 'anomaly free', and the rules to enhance the future prediction of the model are updated. If the recovery process does not succeed, a new weight is assigned.

We specify the recovery mechanism using the following aspects: **Analysis**: relies on the current or historic observation. **Observation**: indicates the type of observed failure (e.g., low response time). **Anomaly**: reflects the kind of fault (e.g., overload). **Reason**: explains the root causes of the problem. **Remedial Action**: explains the solution that can be applied to solve the problem. **Requirements**: steps and constraints that should be considered to apply the action(s). We apply this two sample strategies below.

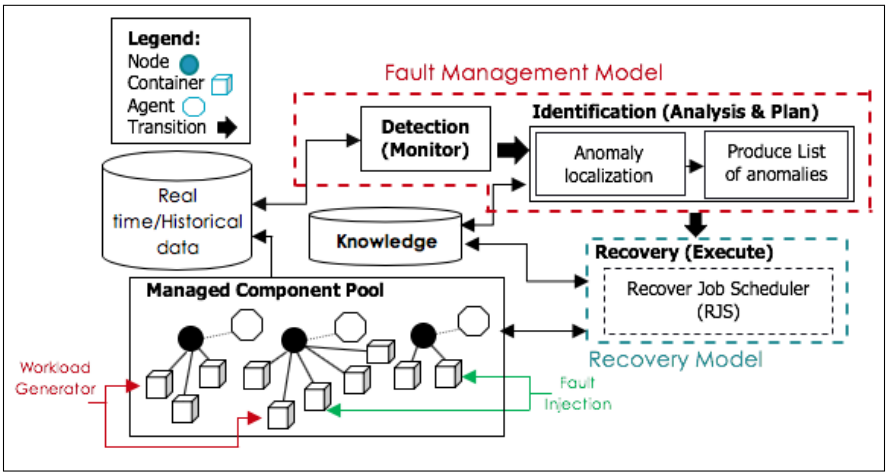


FIGURE 1. THE PROPOSED FAULT MANAGEMENT FRAMEWORK.

D. Motivating Failure/Fault Cases and Recovery Strategies

In the following, we present two samples failure-fault situations, and suitable recovery strategies. The recovery strategies are applied based on the observed response time (current and historic observations), and its related hidden fault states. We illustrate two sample cases–overloaded neighbouring container, and node overload.

1) Container Neighbour Overload (external dependency): this happens when a container c_3 in node N_2 is linked to another container c_2 in another node N_1. In another case, some containers c_3 and c_4 in N_2 dependent on each other, and container c_2 in N_1 depends on c_3. In both cases c_2 in N_1 is badly affected once c_3 or c_4 in N_2 are heavily loaded. This results in a low response time observed from those containers.
Analysis: based on the current/historic observations, hidden states
Observation: low response time at the connected containers (overall failure to meet performance targets).
Anomaly: overload in one or more containers results in underload for another container at different node.
Reason: heavily loaded container with external dependent one (communication)
Remedial Actions:

Option 1: Separate the overloaded container and the external one depending on it from their nodes. Then, create a new node containing the separated containers considering the cluster capacity. Redirect other containers that communicate with these 2 containers in the new node. Connect current nodes with the new one, and calculate the probability of the whole model to know the number of transitions (to avoid the occurrence of overload), and to predict the future behaviour.

Option 2: For the anomalous container, add a new one to the node that has the anomalous container to provide fair workload distribution among containers considering the node resource limits. Or, if the node does not yet reach the resource limits available, move the overloaded container to another node with free resource limits. At the end, update the node.

Option 3: create another node within the node with anomalous container behaviour. Next, direct the communication of current containers to this node. We need to redetermine the probability of the whole model to redistribute the load between containers. Finally, update the cluster and the nodes.

Option 4: distribute load.

Option 5: rescale node.

Option 6: do nothing, if the observed failure relates to a regular system maintenance/update, then no recovery is applied.
Requirements: need to consider node capacity.

2) Node overload (self-dependency)

Analysis: current and historic observations
Observation: low response time at node level (a failure).
Anomaly: overloaded node.
Reason: limited node capacity.
Remedial Actions: *Option 1*: distribute load. *Option 2*: rescale node. *Option 3*: do nothing.
Requirements: collect information regarding containers and nodes, consider node capacity and rescale node(s).

IV. ANOMALY DETECTION AND ANALYSIS

A failure is the inability of a system to perform its required functions within specified performance requirements. Faults (or anomalies) describe an exceptional condition occurring in the system operation that may cause one or more failures. It is a manifestation of an error in system [24]. We assume that a failure is an undesired response time observed during a system component runtime (i.e., observation). For example, fluctuations in workload are faults that may cause a slowdown in system response time (observed failure).

A. Motivation

As an example, Figure 2 shows several observed failures and related resource faults in a test environment. These failures occurred either at a specific time (e.g., F_1, F_9) or over a period

of time (e.g., $F_2 - F_8$). These failures result from fluctuations in resource utilization (e.g., CPU). Utilization measures a resources capacity that is in use. It helps us in knowing the resource workload, and helps us in reducing the amount of jobs from the overloaded resources, e.g., a resource is saturated when its usage is over 50% of its maximum capacity.

The response time varies between high, low and normal categories. It is associated with (or caused by) resource workload fluctuations (e.g., overload, underload or normal load). The fluctuations in workload shall be categorised into states that reflect faults. The anomalous response time is the observed failure that we use initially to identify the type of workload that causes the anomalies. In more concrete terms, we can classify the response time by the severity of a usage anomaly on a resource: low response time (L) varies from $501 - 1000ms$, normal response time (N) reflects the normal operation time of a resource and varies from $201 - 500ms$, and high response time (H) occurs when a response time is less than or equal to $200ms$, which can be used to transfer the workload from the heavy loaded resources to the underloaded resources.

As a result, the recovery strategy differs based on the type of observed failure and the hidden fault. The period of recovery, which is the amount of time taken to recover, differs based on: (1) the number of observed failures, (2) the volume of transferred data (nodes with many tasks require longer recovery time), and (3) network capacity.

B. Observed Failure to Fault Mapping

The first problem is the association of underlying hidden faults to the observed failures. For the chosen metrics (e.g., resource utilization, response time), we can assume prior knowledge regarding (1) the dependency between containers, nodes and clusters; (2) past response time fluctuations for the executable containers; and (3) workload fluctuations that cause changes in response time. These can help us in identifying the mapping between anomalies and failures. An additional difficulty is the hierarchical organisation of clusters consisting of nodes, which themselves consist of containers. We associate an observed container response time to its cause at container, node, or cluster level, where for instance also a neighbouring container can cause a container to slow down. We define a mapping based on an analysis of possible scenarios.

The interaction between the cluster, node and container components in our architecture is based on the following assumptions. A cluster, which is the root node, is composed of multiple nodes, and it is responsible for managing the nodes. A node, which is a virtual machine, has a capacity (e.g., resources available on the node such as memory or CPU). The main job of the node is to submit requests to its underlying substates (containers). Containers are self-contained, executable software packages. Multiple containers can run on the same node, and share the operating environment with other containers. Observations include the emission of failure from a state (e.g., high, low, or normal response time may emit from one or more states). Observation probabilities express the probability of an observation being generated from a resource state. We need to estimate the observation probabilities in order to know under which workloads large response time fluctuations occur and therefore to efficiently utilize a system resource while achieving good performance.

We need a mechanism that dynamically detects the type of anomaly and identifies its causes using this mapping. We identify different cases that may occur at container, node or cluster levels as illustrated in Figure 3. These detected cases serve as a mapping between observable and hidden states, each annotated with a probability of occurrence that can be learned from a running system as a cause will often not be identifiable with certainty.

1) Low Response Time Observed at Container Level: There are different reasons that may cause this:

- *Case 1.1. Container overload (self-dependency)*: means that a container is busy, causing low response times, e.g., c_1 in N_1 has entered into load loop as it tries to execute its processes while N_1 keeps sending requests to it, ignoring its limited capacity.
- *Case 1.2. Container sibling overloaded (internal container dependency)*: this indicates another container c_2 in N_1 is overloaded. This overloaded container indirectly affects the other container c_1 as there is a communication between them. For example, c_2 has an application that almost consumes its whole resource operation. The container has a communication with c_1. At such situation, when c_2 is overloaded, c_1 goes into underload, because c_2 and c_1 share the resources of the same node.
- *Case 1.3. Container neighbour overload (external container dependency)*: this happens when a container c_3 in N_2 is linked to another container c_2 in another node N_1. In another case, some containers c_3, and c_4 in N_2 dependent on each other and container c_2 in N_1 depends on c_3. In both cases c_2 in N_1 is badly affected once c_3 or c_4 in N_2 are heavily loaded. This results in low response time observed from those containers.

2) Low Response Time Observed at Node Level: There are different reasons that cause such observations:

- *Case 2.1. Node overload (self-dependency)*: generally node overload happens when a node has low capacity, many jobs waited to be processed, or when there is a problem in network. Example, N_2 has entered into self load due to its limited capacity, which causes an overload at the container level as well c_3 and c_4.
- *Case 2.2. External node dependency*: occurs when a low response time is observed at node neighbour level, e.g., when N_2 is overloaded due to low capacity or network problem, and N_1 depends on N_2. Such overload may cause low response time observed at the node level, which slows the whole operation of a cluster because of the communication between the two nodes. The reason behind that is N_1 and N_2 share the resources of the same

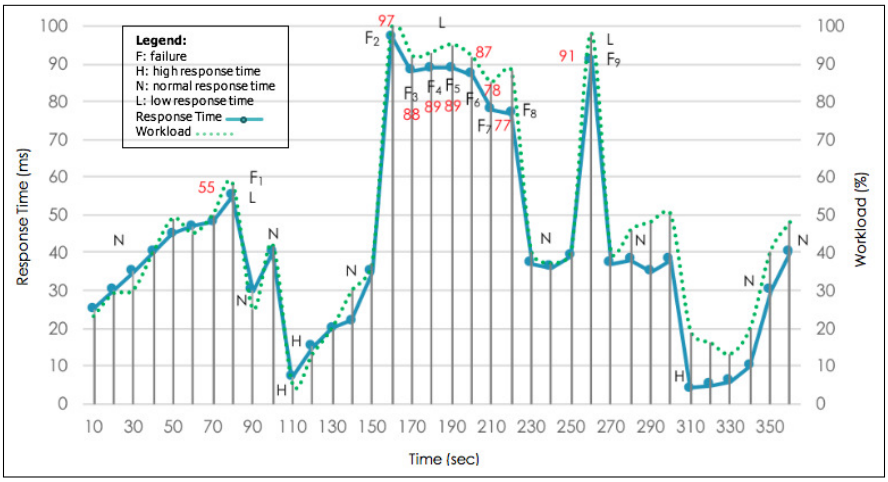


FIGURE 2. RESPONSE TIME AND WORKLOAD FLUCTUATIONS.

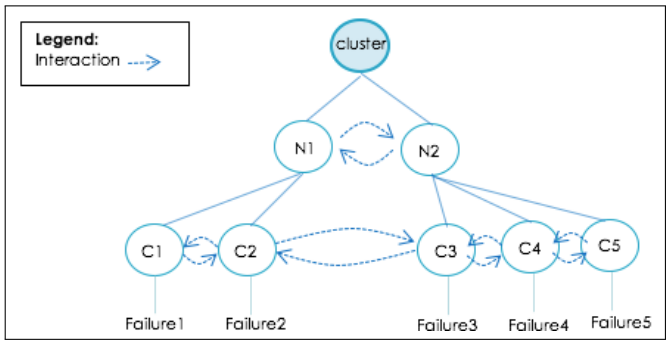


FIGURE 3. THE INTERACTION BETWEEN CLUSTER, NODES AND CONTAINER.

cluster. Thus, when N_1 shows a heavier load, it would affect the performance of N_2.

3) Low Response Time Observed at Cluster Level (Cluster Dependency): If a cluster coordinates between all nodes and containers, we may observe low response time at container and node levels that cause difficulty at the whole cluster level, e.g., nodes disconnected or insufficient resources.

- *Case 3.1. Communication disconnection* may happen due to problem in the node configuration, e.g., when a node in the cluster is stopped or disconnected due to failure or a user disconnect.
- *Case 3.2. Resource limitation* happens if we create a cluster with too low capacity which causes low response time observed at the system level.

This mapping between anomalies and failures across the three hierarchy layers of the architecture needs to be formalised in a model that distinguishes observations and hidden states, and that allows weight to be attached. Thus, the HHMMs are used to reflect the system topology.

C. Hierarchical Hidden Markov Model

The Hierarchical Hidden Markov Model (HHMM) is a generalization of the Hidden Markov Model (HMM) that is used to model domains with hierarchical structure (e.g., intrusion detection, plan recognition, visual action recognition). The HHMM can characterize the dependency of the workload (e.g., when at least one of the states is heavy loaded). The states (cluster, node, container) in the HHMM are hidden from the observer, and only the observation space is visible (response time). The states of the HHMM emit sequences rather than a single observation by a recursive activation of one of the substates (nodes) of a state (cluster). This substate might also be hierarchically composed of substates (containers). Each container has an application that runs on it. In case a node or a container emit observation, it is considered a production state. The states that do not emit observations directly are called internal states. The activation of a substate by an internal state is a vertical transition that reflects the dependency between states. The states at the same level have horizontal transitions. Once the transition reaches to the End state, the control returns to the root state of the chain as shown in Figure 4. The edge direction indicates the dependency between states.

The HHMM is identified by $HHMM =< \lambda, \theta, \pi >$. The λ is a set of parameters consisting of horizontal ζ and vertical χ transitions between states q^d, state transition probability A, observation probability distribution B, initial transition π; d specifies the number of vertical levels, i the horizontal level index, the state space SP at each level and the hierarchical parent-child relationship q_i^d, q_i^{d+1}. The Σ consists of all possible observations O. γ_{in} is the transition to q_j^d from any q_i^d. γ_{out} is the transition of leaving q_j^d from any q_i^d.

We choose HHMM as every state can be represented as a multi-levels HMM in order to:

1) show communication between nodes and containers,
2) demonstrate impact of workloads on the resources,
3) track the anomaly cause,
4) represent the response time variations that emit from nodes and containers.

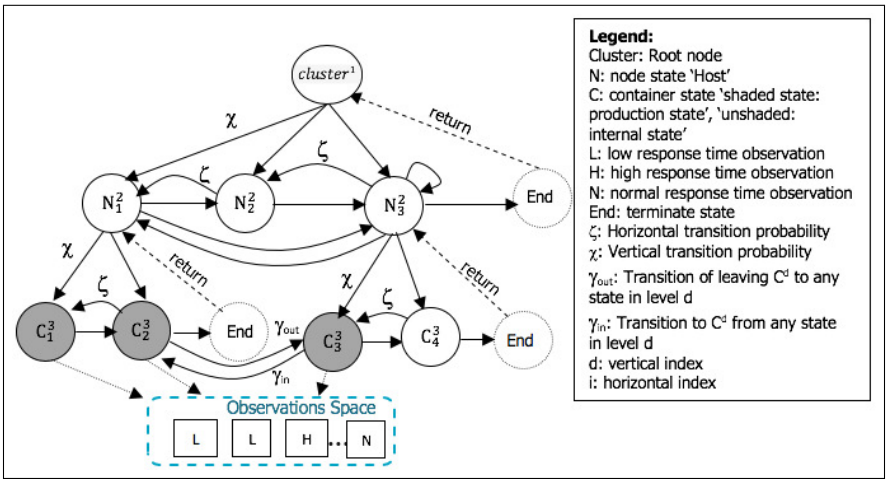


FIGURE 4. HHMM FOR WORKLOAD.

D. Detection and Root Cause Identification using HHMM

Each state may show an overload, underload or normal load state. Each workload is correlated to the resource utilization such as the CPU, and it is associated with the response time observations that are emitted from a container or node through the above case mapping. The existence of anomalous workload in one state not only affects the current state, but it may also affect the other states in the same level or across the levels. The vertical transitions in Figure 4 trace the fault and identify the fault-failures relation. The horizontal transitions show the request/reply transfered between states.

The observation O is denoted by $F_i = \{f_1, f_2, ..., f_n\}$ to refer to the response time observations sequence (failures). The substate and production states are denoted by N and C respectively. A node space SP containing a set of containers, $N_1^2 = \{C_1^3, C_2^3\}$, $N_3^2 = \{C_3^3, C_4^3\}$. Each container produces an observation that reflects the response time fluctuation, $C_1^3 = \{f_1\}, C_2^3 = \{f_1\}, C_3^3 = \{f_2\}$. A state C starts operation at time t if the observation sequence $(f_1, f_2, ..., f_{n-1})$ is generated before the activation of its parent state N. A state ends its operation at time t if the F_t is the last observation generated by any of the production states C reached from N, and the control returns to N from C_{end}. The state transition probability $A_{ij}^{N_i^d} = (a_{ij}^{N^d})$, $a_{ij}^{N^d} = P(N_j^{d+1}|N_i^{d+1})$ indicates the probability of making a horizontal transition from N_i^d to N_j^d. Both states are substates of $cluster^1$.

An observed low response time might reflect some overload (OL). This overload can occur for a period of time or at a specific time before the state might return to the normal load (NL) or underload (UL). This fluctuation in workload is associated with a probability that reflects the state transition status from the OL to NL ($PF_{OL \to NL}$) at a failure rate $\Re$, which indicates the number of failures for a N, C or $cluster$ over a period of time. Sometimes, a system resource remains OL/UL without returning to its NL. We reflect this type of fault as a self-transition overload/underload with probability PF_{OL} (PF_{UL}). Further, a self-transition is applied on normal load PF_{NL} to refer to continuous normal behaviour. In order to address the reliability of the proposed fault analysis, we define a fault rate based on the number of faults occurring during system execution $\Re(FN)$ and the length of failure occurrences $\Re(FL)$ as depicted in "(1)" and "(2)".

$$\Re(FN) = \frac{No\ of\ Detected\ Faults}{Total\ No\ of\ Faults\ of\ Resource} \quad (1)$$

$$\Re(FL) = \frac{Total\ Time\ of\ Observed\ Failures}{Total\ Time\ of\ Execution\ of\ Resource} \quad (2)$$

As failure varies over different periods of time, we can also determine the $Average\ Failure\ Length\ (AFL)$. These metrics feed later into a proactive recovery mechanism. Possible observable events can be linked to each state (e.g., low response time may occur for an overload state or normal load) to determine the likely number of failures observed for each state, and to estimate the total failures numbers for all the states. To estimate the probability of a sequence of failures (e.g., probability of observing low response time for a given state). Its sum is based on the probabilities of all failure sequences that generated by (q^{d-1}), and where (q_i^d) is the last node activated by (q^{d-1}) and ending at the End state. This is done by moving vertically and horizontally through the model to detect faulty states. Once the model reaches the end state, it has recursively moved upward until it reaches the state that triggered the substates. Then, we sum all possible starting states called by the $cluster$ and estimate the probability.

We use the generalized Baum-Welch algorithm [10] to train the model by calculating the probabilities of the model parameters. As shown in "(3)" and "(4)", first, we calculate the number of horizontal transitions from a state to another, which are substates from q^{d-1}, using ξ as depicted in "(3)". The γ_{in} refers to the probability that the O is started to be emitted for $state_i^d$ at t. $state_i^d$ refers to container, node, or cluster. The γ_{out} refers to the O of $state_i^d$ are emitted and

finished at t. Second, as in "(4)", $\chi(t, C_i^d, N_l)$ is calculated to obtain the probability that $state^{d-1}$ is entered at t before O_t to activate state $state_i^d$. The α and β denote the forward and backward transition from bottom-up.

$$\xi(t, C_i^d, C_{End}^d, N_l) = \frac{1}{P(O|\lambda)} \left[\sum_{s=1}^{t} \gamma_{in}(N_l, cluster)\ \alpha(t, C_i^d, N_l)\right] a_{End}^{C_l} \gamma_{out}(t, C_l, cluster) \tag{3}$$

$$\chi(t, C_i^d, N_l) = \frac{\gamma_{in}(t, N_l, cluster)\pi^{N_l}(C_i^d)}{P(O|\lambda)} \left[\sum_{e=t}^{T} \beta(t, e, C_i^d, N_l)\gamma_{out}(e, N_l, cluster)\right] \tag{4}$$

The output of the algorithm is used to train the Viterbi algorithm to find the anomalous hierarchy of the detected anomalous states. As shown in "(5)-(7)", we recursively calculate $\Im$ which is the ψ for a time set $(\bar{t} = \psi(t, t+k, C_i^d, C^{d-1}))$, where ψ is a state list, which is the index of the most probable production state to be activated by C^{d-1} before activating C_i^d. $\bar{t}$ is the time when C_i^d is activated by C^{d-1}. The δ is the likelihood of the most probable state sequence generating $(O_t, \cdots, O_{(t+k)})$ by a recursive activation. The τ is the transition time at which C_i^d is called by C^{d-1}. Once all the recursive transitions are finished and returned to $cluster$, we get the most probable hierarchies starting from $cluster$ to the production states at T period through scanning the sate list ψ, the states likelihood δ, and transition time τ.

$$L = \max_{(1 \le r \le N_i^d)} \left\{\delta(\bar{t}, t+k, N_r^{d+1}, N_i^d)\, a_{End}^{N_i^d}\right\} \tag{5}$$

$$\Im = \max_{(1 \le y \le N^{j-1})} \left\{\delta(t, \bar{t}-1, N_i^d, N^{d-1}) a_{End}^{N^{d-1}}\, L\right\} \tag{6}$$

$$stSeq = \max_{cluster} \left\{\delta(T, cluster), \tau(T, cluster), \psi(T, cluster)\right\} \tag{7}$$

Once we have trained the model, we compare the detected hierarchy against the observed one to detect and identify the type of workload. If the observed hierarchy and detected one is similar, and within the specified threshold, then the status of the observed component is declared as 'Anomaly Free', and the architecture returns to gather more data for further investigation. Otherwise, the hierarchy with the lowest probabilities is considered anomaly. Once we detect and identify the workload type (e.g., OL), a path of faulty states (e.g., $cluster$, N_1^2, C_2^3 and C_3^3) is obtained that reflects observed failures. We repeat these steps until the probability of the model states become fixed. Each state is correlated with time that indicates: the time of its activation, its activated substates, and the time at which the control returns to the calling state. This helps us in the recovery procedure as the anomalous state is recovered first come-first heal.

E. Workload and Resource Utilization Correlation

To check if the occurrence of an anomaly at cluster, node, container resource due to a workload, we calculate the correlation between the workload (user transactions), and resource utilization to specify thresholds for each resource. The user transactions refer to the request rate per second. Thus, we used the Spearman's rank correlation coefficient to generate threshold to indicate the occurrence of fault at the monitored metric in multiple layers.

Our target is to group similar workload for all containers that run the same application in the same period. So that the workloads in the same period have the similar user transactions and resource demand. We add a unique workload identifier to the group of workloads in the same period to achieve traceability through the entire system. We utilize the probabilities of states transitions that we obtain from the HHMM to describe workload during T period. We transform the obtained probabilities to get a workload behaviour vector ω to characterize user transactions behaviours as in "(8)".

$$\omega = \{C_{i=1}^{d=3}, \cdots, C_{j=m}^{d=n}, \cdots, N_{i=1}^{d=2}, \cdots, N_{j=m}^{d=n}, \cdots, cluster\} \tag{8}$$

The correlation between the workload and resource utilization metric is calculated in the normal load behaviour to be a baseline. In case the correlation breaks down, then this refers to the existence of anomalous behaviour (e.g., OL).

V. Evaluation

The proposed architecture is run on the Kubernetes and docker containers. We deploy the TPC-W[1] benchmark on the containers to validate the architecture. We focus on three types of faults the CPU hog, Network packet loss/latency, and performance anomaly caused by workload congestion.

A. Environment Set-Up

To evaluate the effectiveness of the proposed architecture, the experiment environment consists of three VMs. Each VM is equipped with Linux OS, 3 VCPU, 2 GB VRAM, Xen 4.11 [2], and an agent. Agents are installed on each VM to collect the monitoring data from the system (e.g., host metrics, container, performance metrics, and workloads), and send them to the storage to be processed. The VMs are connected through a 100 Mbps network. For each VM, we deploy two containers, and we run into them the TPC-W benchmark.

The TPC-W benchmark is used for resource provisioning, scalability, and capacity planning for e-commerce websites. The TPC-W emulates an online bookstore that consists of 3 tiers: client application, web server, and database. Each tier is installed on VM. We do not consider the database tier in the anomaly detection and identification, as a powerful VM should be dedicated to the database. The CPU and Memory utilization

[1] http://www.tpc.org/tpcw/

[2] https://xenproject.org/

are gathered from the web server, while the Response time is measured from clients end. We run the TPC-W for 300 min. The number of records that we obtained from the TPC-W is 2000.

We use the docker $stats$ command to obtain a live data stream for running containers. The SignalFX Smart Agent[3] monitoring tool is used and configured to observe the runtime performance of components and their resources. We also use the Heapster[4] to group the collected data, and store them in a time series database using the InfluxDB[5]. The data from the monitoring and from datasets are stored in the Real-Time/Historical Data storage to enhance the future anomaly detection. The gathered datasets are classified into training and testing datasets 50% for each. The model training lasts 150 minutes.

B. Fault Scenarios

To simulate real anomalies of the system, script is written to inject different types of anomalies into nodes and containers. The anomaly injection for each component last 5 minutes to be in total 30 minutes for all the system components. The starting and end time of each anomaly is logged.

- CPU Hog: such anomaly is injected to consume all the CPU cycles by employing infinite loops. The stress[6] tool is used to create pressure on CPU
- Network packet loss/latency: the components are injected with anomalies to send or accept a large amount of requests in network. Pumba[7] is used to cause network latency and package loss
- Workload contention: web server is emulated using client application, which generates workload (using Remote Browser Emulator) by simulating a number of user requests that is increased iteratively. Since the workload is always described by the access behaviour, we consider the container is gradually loaded within [30-2000] emulated users requests, and the number of requests is changed periodically. The client application reports response time metric, and the web server reports CPU and Memory utilization. To measure the number of requests and response (latency), the HTTPing[8] is installed on each node. Also, the AWS X-Ray[9] is used to trace of the request through the system.

C. Fault-Failure Mapping Detection and Identification

To address the fault-failure cases, the fault injection (CPU Hog and Network packet loss/latency) is done at two phases: (1) the system level (nodes), (2) components such as nodes and containers, one component at a time. The detection and identification are different as the injection time is varied from one component to another. The injection pause time between each injected fault is 180 sec.

a) Low Response Time Observed at Container Level: Case 1.1. Container overload (self-dependency): here, we add a new container C_5^3 in N_1^2, and we inject it by one anomaly at a time. For the CPU Hog, the anomaly is injected at 910 sec. It takes from the model 30 sec to detect the anomaly and 15 sec to localize it. For the Network packet loss/latency, the injection of anomaly happens at 1135 sec, and the model detects and identifies the anomaly at 1145 and 1163 sec respectively as shown in Table I.

TABLE I. CONTAINER OVERLOAD SELF-DEPENDENCY ANOMALY SCENARIO.

Container overload				
Anomaly				
Injection		Detection	Localization	Type
C^3_5	910	940	955	CPU hog
	1135	1145	1163	Network

Case 1.2. Container sibling overloaded (internal container dependency): in this case, the injection occurs at C_3^3 which in relation with C_4^3. The CPU injection begin at 700 sec for C_3^3, the model detects the anomalous behaviour at 710 sec and localizes it at 725 sec. For Network packet loss/latency, the injection of anomaly occurs at 905 sec. The model needs 46 sec for the detection and 19 sec for the identification. For the C_4^3 the detection happens 34 sec later the detection of C_3^3 for the CPU Hog and the anomaly is identified at 754 sec. For the Network, the detection and identification occur at 903 and 990 sec respectively as shown in Table II.

TABLE II. CONTAINER OVERLOAD INTERNAL-DEPENDENCY ANOMALY SCENARIO.

Container overload				
Anomaly				
Injection		Detection	Localization	Type
C^3_3	700	710	725	CPU hog
	905	951	970	Network
C^3_4		744	754	CPU hog
		903	990	Network

Case 1.3. Container neighbour overload (external container dependency): at this case, a CPU Hog is injected at C_1^3 which in relation with C_3^3. The injection begin at 210 sec. After training the HHMM, the model detects and localizes the anomalous behaviour for C_1^3 at 225 and 230 sec. For Network fault, the injection occurs at 415 sec for C_1^3. The model takes 429 sec for the detection and 450 sec for the identification. While for C_3^3, the CPU and Network faults are detected at 215/423 sec and identified at 240/429 sec as shown in Table III.

b) Low Response Time Observed at Node Level: Case 2.1. Node overload (self-dependency): at this case we create a new node N_4^2 with small application and we inject the node by one anomaly at a time. For the CPU Hog, the anomaly is injected at N_4^2. The injection begins at 413 sec. After training the HHMM, the model detects the anomalous behaviour at

[3]https://www.signalfx.com/
[4]https://github.com/kubernetes-retired/heapster
[5]https://www.influxdata.com/
[6]https://linux.die.net/man/1/stress
[7]https://alexei-led.github.io/post/pumba_docker_netem/
[8]https://www.vanheusden.com/httping/
[9]https://aws.amazon.com/xray/

TABLE III. CONTAINER OVERLOAD EXTERNAL-DEPENDENCY ANOMALY SCENARIO.

Container overload				
Anomaly				
Injection		Detection	Localization	Type
C^3_1	210	225	230	CPU hog
	415	429	450	Network
C^3_3		215	240	CPU hog
		423	429	Network

443 sec and localizes it at 461 sec. For the Network packet loss/latency, the injection of anomaly happens at 1210 sec, and the model detects and identifies anomaly at 1260 and 1275 sec respectively as shown in Table IV.

TABLE IV. NODE OVERLOAD SELF-DEPENDENCY ANOMALY SCENARIO.

Node overload				
Anomaly				
Injection		Detection	Localization	Type
N^2_4	413	443	461	CPU hog
	1210	1260	1275	Network

Case 2.2. External node dependency: at such situation, a CPU Hog anomaly is injected at N^2_1. The injection begins at 813 sec. After training the HHMM, the model detects the anomalous behaviour at 846 sec and localizes it at 862 sec. For Network packet loss/latency, the injection of anomaly occurs at 1024 sec. The model needs 1084 sec for the detection and 1115 sec for the identification as shown in Table V.

TABLE V. NODE OVERLOAD EXTERNAL-DEPENDENCY ANOMALY SCENARIO.

Node overload				
Anomaly				
Injection		Detection	Localization	Type
N^2_1	813	846	862	CPU hog
	1024	1084	1115	Network

c) Low Response Time Observed at Cluster Level (Cluster Dependency): Case 3.1. Communication disconnection: at this case, we terminate the containers in N^2_3, and we send a request to the TPC-W server (N^2_3). The detection and identification for each network fault are 585 sec for the detection and 610 sec for the identification as shown in Table VI.

TABLE VI. CLUSTER OVERLOAD COMMUNICATION DISCONNECTION ANOMALY SCENARIO.

Cluster overload		
Anomaly		
	Detection	Localization
N^2_3	585	610

Case 3.2. Resource limitation: at this case, we inject N^2_1, and N^2_3 at the same time with the CPU Hog fault to exhaustive the nodes capacity. The injection, detection, and identification are 1120, 1181, and 1192 sec. For the Network fault, the injection happens at 1372 sec, and the detection, and identification are at 1387, and 1392 sec as shown in Table VII.

TABLE VII. CLUSTER OVERLOAD RESOURCE LIMITATION ANOMALY SCENARIO.

Cluster overload				
Anomaly				
Injection		Detection	Localization	Type
N	1120	1181	1192	CPU hog
	1372	1387	1392	Network

D. Detection and Identification of Workload Contention

For the workload, to show the influence of workload on CPU utilization monitored metric, we measure the response time (i.e., the time required to process requests), and throughput (i.e., the number of transactions processed during a period). We first generate gradual requests/sec at the container level. The number of user requests increases from 30 to 2000 with a pace of 10 users incrementally, and each workload lasts for 10 min. As shown in Figure 5, the results show that the throughput increases when the number of requests increases, then it remains constant once the number of requests reaches 220 request/sec. This means that when the number of user requests is reached 220 request/sec, the utilization of CPU reaches a bottleneck at 90%, and the performance degrades. On the other hand, the response time keep increasing with the increasing number of requests as shown in Figure 6. The result demonstrated that the dynamic workloads have a noticeable impact on the container metrics as the monitored containers are unable to process more than those requests. We also notice that there is a linear relationship between the number of concurrent users and the CPU utilization before resource contention in each user transaction behaviour pattern. We calculate the correlation between the monitored metric, and the number of user requests. We obtain a strong correlation between the two measured variables reaches 0.25775 for two variables. The result concludes that the number of requests influences the performance of the monitored metrics.

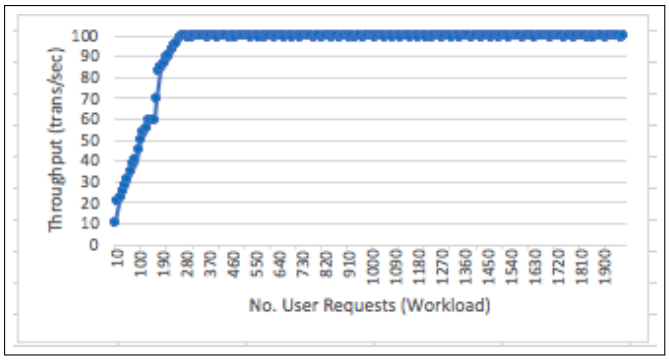


FIGURE 5. WORKLOAD - THROUGHPUT AND NUMBER OF USER REQUESTS.

E. Assessment of Detection and Identification

The model performance is compared with other techniques such as the Dynamic Bayesian Network (DBN) and the Hierarchical Temporal Memory (HTM). To evaluate the effectiveness

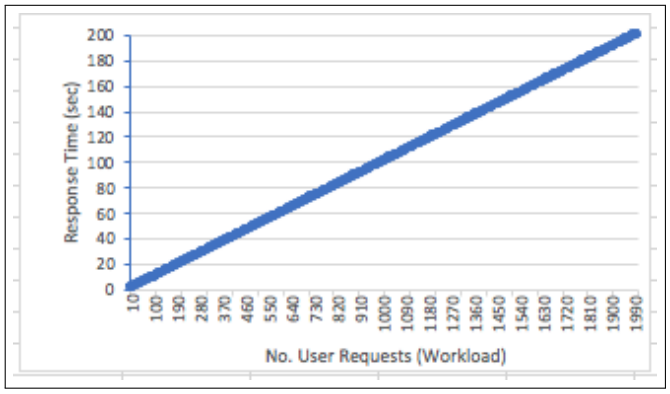


FIGURE 6. WORKLOAD - RESPONSE TIME AND NUMBER OF USER REQUESTS.

of anomaly detection, common measures [25] in anomaly detection are used:

Root Mean Square Error (RMSE) measures the differences between detected and observed value by the model. A smaller RMSE value indicates a more effective detection scheme.

Mean Absolute Percentage Error (MAPE) measures the detection accuracy of a model. Both RMSE and MAPE are negatively-oriented scores, i.e., lower values are better.

Number of Correctly Detected Anomaly (CDA) It measures percentage of the correctly detected anomalies to the total number of detected anomalies in a given dataset. High CDA indicates the model is correctly detected anomalous behaviour.

Recall measures the completeness of the correctly detected anomalies to the total number of anomalies in a given dataset. Higher recall means that fewer anomaly cases are undetected.

Number of Correctly Identified Anomaly (CIA) is the number of correct identified anomalies (NCIA) out of the total set of identification, which is the number of correct identification (NCIA) + the number of incorrect identifications (NICI)). The higher value indicates the model is correctly identified anomalous component.

$$CIA = \frac{NCIA}{NCIA + NICI} \tag{9}$$

Number of Incorrectly Identified Anomaly (IIA) is the number of identified components which represents an anomaly but misidentified as normal by the model. A lower value indicates that the model correctly identified anomalies.

$$IIA = \frac{FN}{FN + TP} \tag{10}$$

FAR is the number of the normal identified component which has been misclassified as anomalous by the model.

$$FAR = \frac{FP}{TN + FP} \tag{11}$$

The false positive (FP) means the detection/identification of anomaly is incorrect as the model detects/identifies the normal behaviour as anomaly. True negative (TN) means the model can correctly detect and identify normal behaviour as normal.

TABLE VIII. VALIDATION RESULTS.

Metrics	HHMM	DBN	HTM
RMSE	0.23	0.31	0.26
MAPE	0.14	0.27	0.16
CDA	96.12%	91.38%	94.64%
Recall	0.94	0.84	0.91
CIA	94.73%	87.67%	93.94%
IIA	4.56%	12.33%	6.07%
FAR	0.12	0.26	0.17

The results in Table VIII depict that both the HHMM and HTM achieve good results for the detection and identification. While the results of the DBN a little bit decayed for the CDA with approximately 5% than the HHMM and 3% than the HTM. The three algorithms can detect obvious anomalies in the datasets. Both the HHMM and HTM show higher detection accuracy as they are able to detect temporal anomalies in the dataset. The result interferes that the HHMM is able to link the observed failure to its hidden workload.

VI. USE CASE DISCUSSION

In order to illustrate the architecture, we discuss here two use cases. The first addresses a widely used cloud setting, where clusters of containers are managed by an orchestration solution such as the Kubernetes. The second looks at an edge cloud scenario, where a cluster of constrained hardware devices hosts container clusters.

A. Use Case: Cloud Container Management

Containers have grown in popularity in recent years and are now widely used as the unit of software deployment, also in cloud environments. Many cloud infrastructure (IaaS) and platform service (PaaS) providers offer container deployment options. In many cases, an orchestration tool like the Kubernetes[10], see Figure 7, that supports automated deployment, scaling and management of containerized applications are used by the providers, see Figures 8 and 7. These are typically homogeneous cloud container cluster in terms of the underlying infrastructure.

A problem that becomes apparent here is that a service consumer have access to monitoring data at the service level, but not necessarily at the underlying (physical) infrastructure level [40]. Nonetheless, service consumer are often given access to controllers that can for instance auto-scale the application deployed.

In this case, the user can be provided with a trained HMM that reflects possible faults for the observed failures.

B. Use Case: Edge Cloud Orchestration

Containers as a more lightweight form of virtualisation compared to virtual machines (VMs) consume less resources. They compare favourably to VMs in terms of startup time to

[10]https://kubernetes.io/

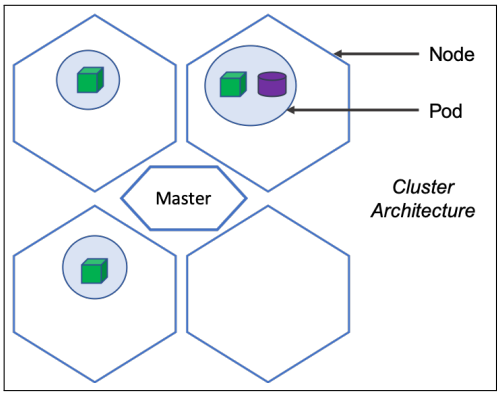


FIGURE 7. A CLUSTER ARCHITECTURE BASED ON KUBERNETES ARCHITECTURAL CONCEPTS.

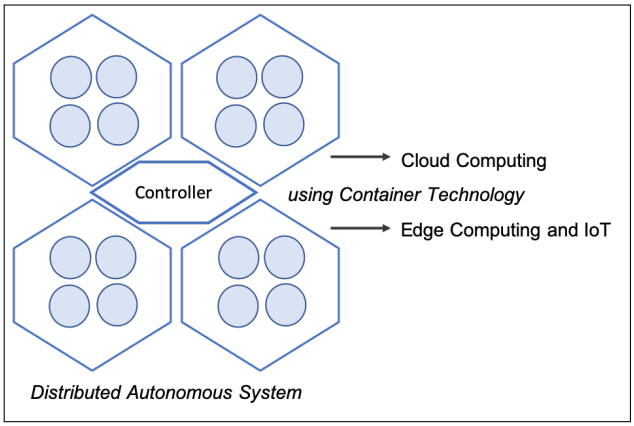


FIGURE 8. A DISTRIBUTED SYSTEM FOR CLOUD AND EDGE COMPUTING BASED ON CONTAINERS.

memory/storage needs. This makes containers more suitable to be utilised outside the classical centralised cloud environment. Here, edge cloud infrastructures that provide computational capabilities for IoT or other remote application can benefit from the containers' lightweightness. This is in particular useful if the edge infrastructure is limited in terms of its capabilities.

For the latter situation, we consider here a cluster on single-board devices as the physical infrastructure to host the

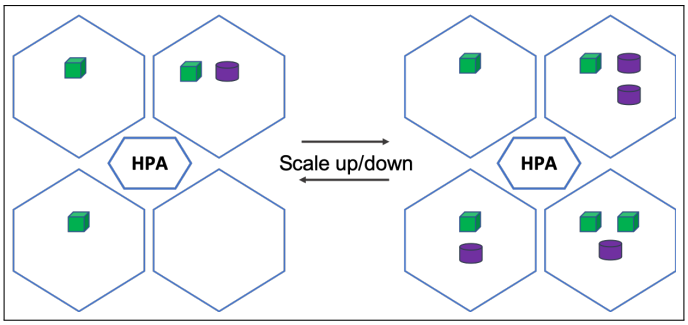


FIGURE 9. KUBERNETES AUTO-SCALING BASED ON THE HPA HORIZONTAL POD AUTOSCALER.

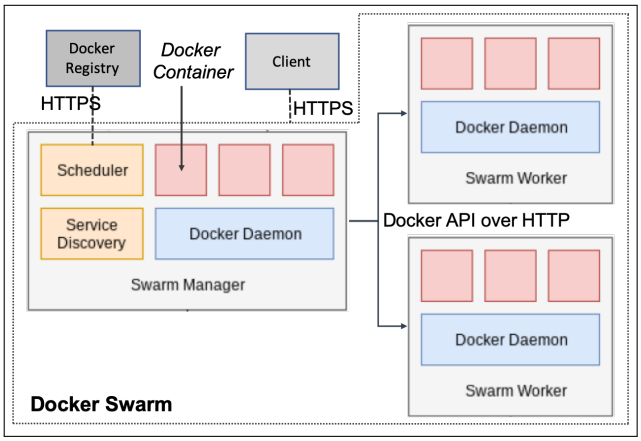


FIGURE 10. A DOCKER SWARM MANAGED ARCHITECTURE FOR CONTAINER ORCHESTRATION.

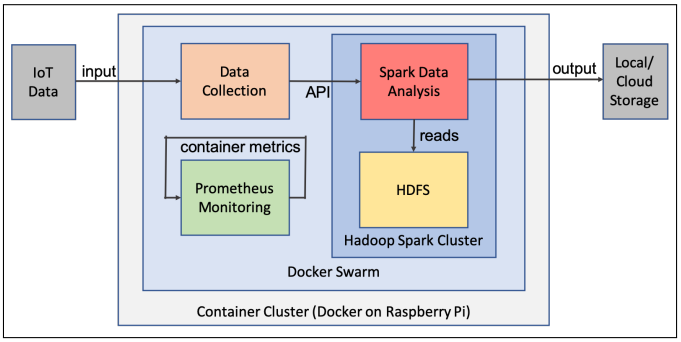


FIGURE 11. A DOCKER CONTAINER ARCHITECTURE FOR DATA STREAM PROCESSING WITH MONITORING SUPPORT.

container cluster platform. Specifically, we use Raspberry Pi[11] devices in this use case. In our experiments, we use the Docker Swarm[12] as the container orchestration tool, see Figure 10.

C. Use Case Scenario: Smart Farming

We categories the fault/failure cases, in which observable failures (to meet QoS requirements) are mapped to their root causes, i.e., the faults that have caused them. Examples are an overloaded container itself or a neighbouring container on which a container depends (e.g., is waiting for an answer) [23], [44]. We use the Markov models to reflect the possibility of several causes and the likelihood of each of these. Typical fault types are the CPU hog, network latency or workload contention.

For each of these mapping cases, we associate suitable remedial actions, such as workload distribution, container migration or resource rescaling.

These can be illustrated in a smart farming scenario. We assume here three central services: an animal stable in which air conditioning and feeding are automated, an outdoor irrigation system and support for tractor and machinery positioning in

[11]https://www.raspberrypi.org/
[12]https://docs.docker.com/engine/swarm/

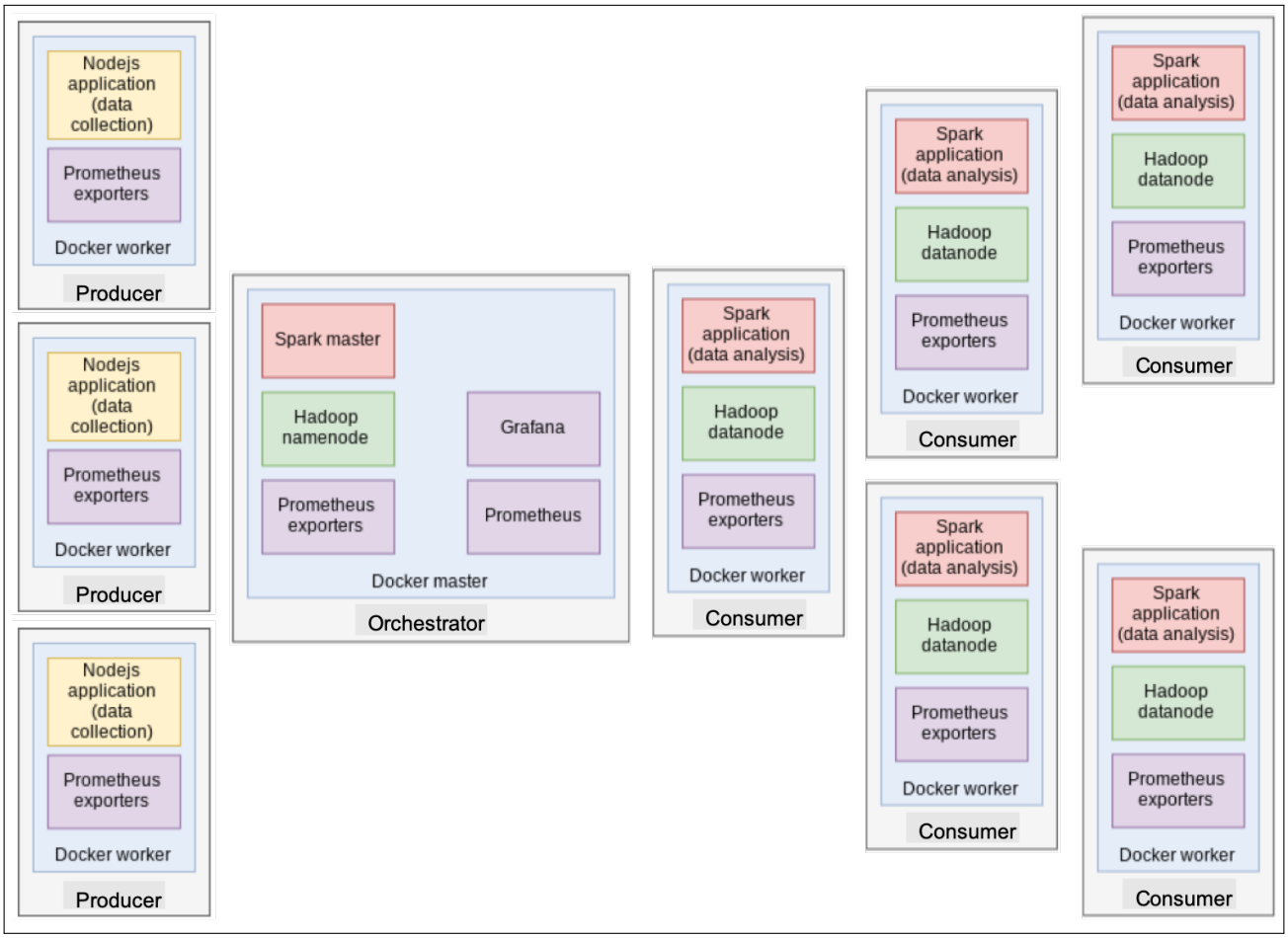


FIGURE 12. A DOCKER CONTAINER DISTRIBUTION FOR A RASPBERRY PI HOSTED CLUSTER.

remote fields. In particular, the outdoor services rely on low-power infrastructure to allow battery and/or solar panel driven energy supply. The indoor service requires reliability based on robust, but also redundant device infrastructures that can work in challenging conditions (e.g., dirt). The following problem situations are possible, and can be supported by our solution:

- Configuration and Testing: during installation or maintenance, increased demands on particular devices (e.g., CPU hogs) can emerge if data-rich test programs are run. Here, migrating containers (i.e., repurposing devices outside the actual service domain) can help.
- Increased Mobility: a higher number of vehicles on the fields might need to be coordinated, for instance during harvest time. This increase the latency problems in the network for both the coordination between vehicles and also recording of data on a central server. Here, moving containers for data preprocessing close to the vehicles can help to reduce the data volume on the network.

Our work in [20], [45] demonstrates how a container cluster solution implemented on Raspberry Pis can support this type of scenario. There, the Docker Swarm based management supports containers for data stream processing (the Apache Park), supported by the Prometheus as a monitoring tool, the Grafana for analyse/visualised data and databases like the InfluxDB to store data, see Figures 11 and 12.

The HMM identifies different anomaly states [1]. These are dependent on the monitored performance and workload/utilisation metrics. In other works [33], [35], we have used fuzzy logic to map monitored data to so-called membership functions that represent these different states. We refer the reader to these works for more detail. Here we focus on the anomaly processing.

VII. DISCUSSION – TRUST ANOMALIES

Anomaly detection and analysis techniques normally address performance and resource management in the context of software systems management. Another quality concern that is different from performance and resource consumption is the context of security and trust. Any open software system has a range of security vulnerabilities. Thus, checking continuously for anomalies in order to find unusual behaviour that might indicate attacks or the loss of information in some form is consequently also a relevant anomaly detection concern.

The concept of trust is here a related aspect that covers security but also the trust into the measurement and handling

of performance and other technical factors. A trust problem occurs if providers and consumers of services meet in an environment where no prior trust relationship exists. A trust anomaly here is a situation in which the delivery of a previously guaranteed service (or the promise of its delivered quality) is in doubt. An anomaly detection solution as the one presented here can help to proactively invoke a remedial action or to record more detailed information (in a tamper-proof way to avoid trust issues to arise). This would allow the analysis and resolution of disputes at a later stage.

The management of trust regarding the Quality-of-Service (QoS) compliance using a trust anomaly management solution shall now be discussed. If a-priori trust does not exist, it is crucial to capture, collect and store necessary information in a tamper-proof way that neither party can interfere with. Distributed ledger technology in the form of blockchains as mechanism to manage anomalies is a possible solution. A blockchain is a distributed data store for digital transactions, ressembling a ledger [55], [48], [48]. The blockchains are used for various applications [59], [63], [58], [56], [57]. These blocks are connected and secured using cryptographic mechanisms. Each of the blocks contains a cryptographic hash of the previous block, and also a timestamp and transaction data. Thus, a blockchain is inherently tamper-proof by design, which means it is resistant to modification of stored data.

This blockchain idea applies in case a consumer requires trustworthy documentation for instance in failure cases, but these blockchains maybe also always be used if a provider need assurance about having provided as planned or promised in a contract. More concretely, an anomaly detection mechanism as we introduced above can now, if the QoS compliance is for example under threat, switch on blockchain storage [55], [41], [37]. This could be as remedial supportive action for later analysis that can provide the required tamper-proof information for the recovery or dispute solving. This solution remains a part of the future work on our anomaly management architecture. However, this short discussion shows that the architecture presented is not limited to performance concerns and immediate remedial actions only, but that other quality concerns can be considered and long-term disputes over the origin and responsibilities can be solved.

VIII. Conclusion and Future Work

Cloud environments cause separation between providers and consumers. The virtualisation in these contexts does anyway separate the physical view from the logical perspective. Furthermore, only providers have access to the infrastructure, which means that consumers cannot always accurately interpret observed anomalies in application and service behaviour. We have introduced a architecture for the detection and identification of anomalies. The key objective is to provide an analysis feature that maps observable quality concerns onto hierarchical hidden resources in a clustered environment and their operation in order to identify the reason for performance degradations and other anomalies.

As the formal model, so-called the Hidden Hierarchical Markov Models (HHMM) are used to represent the hierarchical nature of the unobservable resources. We have analysed mappings between observations and resource usage based on a clustered container scenario. To evaluate the performance of the proposed architecture, the HHMM is compared with other machine learning algorithms such as the Dynamic Bayesian Network (DBN), and the Hierarchical Temporal Memory (HTM). The results show that the proposed architecture is able to detect and identify anomalous behaviour with more than 96%, which demonstrates the suitability of the solution.

We have been focusing specifically on clustered cloud environments as the architectural setting [52], [49], [69], [68], ultimately aiming at self-adaptation in the recovery process [66], [51]. In addition, we have selected the now widely used container technology as the deployment solution [40], [78]. The use cases that we have discussed here reflect this setting and show the suitability of the proposed architecture in this context.

As part of our future work, we are planning to fully implement the architecture. Also, carrying out further experiments is expected to fully confirm these given conclusions here is a wider range of application settings. Furthermore, another aim is to provide a self-healing mechanism to recover the localized anomalies detected.

On a more practical side, we want to follow the focus on containers further and aim to explore concerns from microservice architectures [21], [65], [67] and containers [32], [20] as their deployment technology in future investigations.

Anomalies are generally considered to be situations that impact on clearly specified system requirements, like performance. These might in turn impact on the user to fulfill her/his objectives with the system in question. An interesting possible investigation in the future could approach this more clearly from the user perspective. Providing a semantic context of activities would here be a first step [62], [53]. As a concrete application area where the user objectives are complex is educational technology systems [71], [72], [73], [74], where learning activities as cognitive processes need to be facilitated [75], [76]. This shall be looked at as well.

References

[1] A. Samir and C. Pahl, “Anomaly Detection and Analysis for Clustered Cloud Computing Reliability,” in *The Tenth International Conference on Cloud Computing, GRIDs, and Virtualization*, 110–119. 2019.

[2] C. Pahl, P. Jamshidi, and O. Zimmermann, “Architectural principles for cloud software,” in *ACM Transactions on Internet Technology (TOIT)*, 18 (2), 17. 2018.

[3] C. Pahl, I. Fronza, N. El Ioini, and H. Barzegar, “A Review of Architectural Principles and Patterns for Distributed Mobile Information Systems,” in *14th Intl Conf on Web Information Systems and Technologies*. 2019.

[4] D. von Leon, L. Miori, J. Sanin, N. El Ioini, S. Helmer, and C. Pahl, “A Lightweight Container Middleware for

Edge Cloud Architectures," in *Fog and Edge Computing: Principles and Paradigms*, 145–170. 2019.

[5] X. Chen, C.-D. Lu, and K. Pattabiraman, "Failure Prediction of Jobs in Compute Clouds: A Google Cluster Case Study," in *International Symposium on Software Reliability Engineering, ISSRE*, pp. 167–177. 2014.

[6] G. C. Durelli, M. D. Santambrogio, D. Sciuto, and A. Bonarini, "On the Design of Autonomic Techniques for Runtime Resource Management in Heterogeneous Systems," PhD dissertation, Politecnico di Milano. 2016.

[7] T. Wang, J. Xu, W. Zhang, Z. Gu, and H. Zhong, "Self-adaptive cloud monitoring with online anomaly detection," in *Fut Gen Computer Systems*, 80:89-101. 2018.

[8] P. Jamshidi, A. Sharifloo, C. Pahl, H. Arabnejad, A. Metzger, and G. Estrada, "Fuzzy self-learning controllers for elasticity management in dynamic cloud architectures," in *12th International ACM SIGSOFT Conference on Quality of Software Architectures, QoSA*, 70–79. 2016.

[9] P. Jamshidi, A. Sharifloo, C. Pahl, A. Metzger, and G. Estrada, "Self-learning cloud controllers: Fuzzy q-learning for knowledge evolution," in *Intl Conference on Cloud and Autonomic Computing*. 208-211. 2015.

[10] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden markov model: analysis and applications," in *Machine Learning*, vol. 32, no. 1, 41–62. 1998.

[11] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications," in *IEEE Transactions on Cloud Computing*, 3(4):449–458. 2015.

[12] N. Sorkunlu, V. Chandola, and A. Patra, "Tracking system behaviour from resource usage data," in *Intl Conference on Cluster Computing*, 410–418. 2017.

[13] M. Peiris, J. H. Hill, J. Thelin, S. Bykov, G. Kliot, and C. Konig, "PAD: Performance anomaly detection in multi-server distributed systems," in *International Conf on Cloud Computing, CLOUD*, June, 769–776. 2014.

[14] T. F. Düllmann, "Performance anomaly detection in microservice architectures under continuous change," Master, U Stuttgart, 2016.

[15] C. Pahl, A. Brogi, J. Soldani, and P. Jamshidi, "Cloud container technologies: a state-of-the-art review," in IEEE Transactions on Cloud Computing. 2018.

[16] S. Maurya and K. Ahmad, "Load Balancing in Distributed System using Genetic Algorithm," in *Intl Journal of Engineering and Technology*, 5(2):139–142. 2013.

[17] H. Sukhwani, "A survey of anomaly detection techniques and hidden markov model," in *International Journal of Computer Applications*, vol. 93, no. 18, 26–31. 2014.

[18] R. Heinrich, A. van Hoorn, H. Knoche, F. Li, L. E. Lwakatare, C. Pahl, S. Schulte, and J. Wettinger, "Performance engineering for microservices: research challenges and directions," in *ACM/SPEC International Conference on Performance Engineering Companion*, 223–226. 2017.

[19] N. Ge, S. Nakajima, and M. Pantel, "Online diagnosis of accidental faults for real-time embedded systems using a hidden Markov model," in *Simulation* 91(19):851-868. 2016.

[20] R. Scolati, I. Fronza, N. El Ioini, A. Samir, and C. Pahl, "A Containerized Big Data Streaming Architecture for Edge Cloud Computing on Clustered Single-Board Devices," in *10th International Conference on Cloud Computing and Services Science*, 68-80. 2019.

[21] D. Taibi, V. Lenarduzzi, and C. Pahl, "Architectural Patterns for Microservices: A Systematic Mapping Study," in *Proceedings CLOSER Conference*, 221–232. 2018.

[22] G. Brogi, "Real-time detection of advanced persistent threats using information flow tracking and hidden markov," Doctoral dissertation. 2018.

[23] A. Samir and C. Pahl, "A Controller Architecture for Anomaly Detection, Root Cause Analysis and Self-Adaptation for Cluster Architectures," in *The Eleventh International Conference on Adaptive and Self-Adaptive Systems and Applications*, 75–83. 2019.

[24] IEEE, "IEEE Standard Classification for Software Anomalies (IEEE 1044 - 2009)," pp. 1–4. 2009.

[25] K. Markham, "Simple guide to confusion matrix terminology," 2014.

[26] B. Magableh and M. Almiani, "A Self Healing Microservices Architecture: A Case Study in Docker Swarm Cluster," in *Intl Conference on Advanced Information Networking and Applications*, 846–858. 2019.

[27] C. Pahl, "An ontology for software component matching," in *International Conference on Fundamental Approaches to Software Engineering*, 6–21. 2003.

[28] A. Khiat, "Cloud-RAIR: A Cloud Redundant Array of Independent Resources," in *Intl Conference on Cloud Computing, Grids, and Virtualization*, 133–137. 2019.

[29] C. Pahl, N. El Ioini, and S. Helmer, "A Decision Framework for Blockchain Platforms for IoT and Edge Computing," in *3rd International Conference on Internet of Things, Big Data and Security*, 105-113. 2018.

[30] M. Hasan, M. Milon Islam, I. Islam, and M. Hashem, "Attack and Anomaly Detection in IoT Sensors in IoT Sites Using Machine Learning Approaches," in *Internet of Things*, vol. 7, 1–14. 2019.

[31] P. Jamshidi, C. Pahl, and N. C. Mendonca, "Pattern-based multi-cloud architecture migration," in *Software: Practice and Experience*, 47 (9), 1159-1184. 2017.

[32] P. Jamshidi, C. Pahl, N. C. Mendonca, J. Lewis, and S. Tilkov, "Microservices: The Journey So Far and Challenges Ahead," in *IEEE Software*, 35 (3), 24-35. 2018.

[33] H. Arabnejad, C. Pahl, P. Jamshidi, and G. Estrada, "A comparison of reinforcement learning techniques for fuzzy cloud auto-scaling," in *17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. 2017.

[34] A. Jindal, V. Podolskiy, and M. Gerndt, "Performance modelling for Cloud Microservice Applications," in *ACM/SPEC International Conference on Performance Engineering*, 25–32. 2019.

[35] P. Jamshidi, C. Pahl, and N. C. Mendonca, "Managing uncertainty in autonomic cloud elasticity controllers," in *IEEE Cloud Computing*, 50–60. 2016.

[36] C. Pahl and B. Lee, "Containers and clusters for edge cloud architectures - A technology review," in *IEEE International Conference on Future Internet of Things and Cloud*, August, 379–386. 2015.

[37] C. Pahl., N. El Ioini, S. Helmer, and B. Lee, "An architecture pattern for trusted orchestration in IoT edge clouds," in *The Third International Conference on Fog and Mobile Edge Computing, FMEC*, April, 63–70. 2018.

[38] C. Sauvanaud, M. Kaâniche, K. Kanoun, K. Lazri, and G. Da Silva Silvestre, "Anomaly detection and diagnosis for cloud services: Practical experiments and lessons learned," in *Jrnl of Systems and Software* 139:84-106. 2018.

[39] L. Mariani, C. Monni, M. Pezze, O. Riganelli, and R. Xin, "Localizing Faults in Cloud Systems," in *IEEE 11th International Conference on Software Testing, Verification and Validation*, 262–273. 2018.

[40] F. Ghirardini, A. Samir, I. Fronza, and C. Pahl, "Performance Engineering for Cloud Cluster Architectures using Model-Driven Simulation," in *ESOCC Workshops - CloudWays'18*. 2019.

[41] C. Pahl, N. El Ioini, S. Helmer, and B. Lee, "A Semantic Pattern for Trusted Orchestration in IoT Edge Clouds," in *Internet Technology Letters*. 2019.

[42] N. Kratzke, "About Microservices, Containers and their Underestimated Impact on Network Performance," in *CoRR*, vol. abs/1710.0. 2017.

[43] T. Zwietasch, "Online Failure Prediction for Microservice Architectures," Master Thesis, U Stuttgart. 2017.

[44] A. Samir and C. Pahl, "Detecting and Predicting Anomalies for Edge Cluster Environments using Hidden Markov Models," in *IEEE International Conference on Fog and Mobile Edge Computing*, 21–28. 2019.

[45] D. Taibi, V. Lenarduzzi, and C. Pahl, "Processes, motivations, and issues for migrating to microservices architectures: An empirical investigation," in *IEEE Cloud Computing*, 4 (5), 22-32. 2017.

[46] O. Ibidunmoye, T. Metsch, and E. Elmroth, "Real-time detection of performance anomalies for cloud services," in *IEEE/ACM Intl Symposium on Quality of Service*. 2016.

[47] A. Wert, "Performance problem diagnostics by systematic experimentation," PhD, KIT. 2015.

[48] N. El Ioini, C. Pahl, and S. Helmer, "A decision framework for blockchain platforms for IoT and edge computing," in *Proceedings of the 3rd International Conference on Internet of Things, Big Data and Security*. 2018.

[49] D. von Leon, L. Miori, J. Sanin, N. El Ioini, S. Helmer, and C. Pahl, "A performance exploration of architectural options for a middleware for decentralised lightweight edge cloud architectures," in *CLOSER Conference*. 2018.

[50] Q. Guan, C. C. Chiu, and S. Fu, "CDA: A cloud dependability analysis framework for characterizing system dependability in cloud computing infrastructures," in *Pacific Rim Intl Symp on Dependable Computing,* 11–20. 2012.

[51] H. Arabnejad, C. Pahl, G. Estrada, A. Samir, and F. Fowley, "A fuzzy load balancer for adaptive fault tolerance management in cloud platforms," in *Europ Conf on Service-Oriented and Cloud Computing*, 109-124. 2017.

[52] P. Jamshidi, C. Pahl, S. Chinenyeze, and X. Liu, "Cloud migration patterns: a multi-cloud service architecture perspective," in *Service-Oriented Computing ICSOC2014 Workshops*, 6-19. 2015.

[53] D. Fang, X. Liu, I. Romdhani, P. Jamshidi, and C. Pahl, "An agility-oriented and fuzziness-embedded semantic model for collaborative cloud service search, retrieval and recommendation," in *Future Generation Computer Systems* 56, 11-26. 2016.

[54] Y. Tan, H. Nguyen, Z. Shen, and X. Gu, "PREPARE : Predictive Performance Anomaly Prevention for Virtualized Cloud Systems," in *Intl Conference on Distributed Computing Systems*, 285–294. 2012.

[55] N. El Ioini and C. Pahl, "Trustworthy Orchestration of Container Based Edge Computing Using Permissioned Blockchain," in *Intl Conference on Internet of Things: Systems, Management and Security*, 147-154. 2018.

[56] C. A. Ardagna, R. Asal, E. Damiani, N. El Ioini, and C. Pahl, "Trustworthy IoT: An Evidence Collection Approach Based on Smart Contracts," in *2019 IEEE International Conference on Services Computing (SCC)*, 46–50. 2019.

[57] V. T. Le, C. Pahl, and N. El Ioini, "Blockchain Based Service Continuity in Mobile Edge Computing," in *6th International Conference on Internet of Things: Systems, Management and Security*, 2019.

[58] C. A. Ardagna, R. Asal, E. Damiani, T. Dimitrakos, N. El Ioini, and C. Pahl, "Certification-based cloud adaptation," in *IEEE Transactions on Services Computing*. 2018.

[59] G. D'Atri, V.T. Le, C. Pahl, and N. El Ioini, "Towards Trustworthy Financial Reports Using Blockchain," in *Proceedings Tenth International Conference on Cloud Computing, GRIDs, and Virtualization*. 2019.

[60] N. El Ioini and C. Pahl, "A Review of Distributed Ledger Technologies," in *On the Move to Meaningful Internet Systems. OTM 2018 Conferences*, 227-288. 2018.

[61] S. Govindan, J. Liu, A. Kansal, and A. Sivasubramaniam, "Cuanta: Quantifying Effects of Shared On-chip Resource Interference for Consolidated Virtual Machines," in *ACM Symp on Cloud Computing*1–14. 2011.

[62] M. Javed, Y. M. Abgaz, and C. Pahl, "Ontology change management and identification of change patterns," in *Journal on Data Semantics* 2(2-3), 119-143. 2013.

[63] S. Helmer, M. Roggia, N. El Ioini, and C. Pahl, "EthernityDB - Integrating Database Functionality into a Blockchain," in *European Conference on Advances in Databases and Information Systems*, 37–44. 2019.

[64] S. Helmer, C. Pahl, J. Sanin, L. Miori, S. Brocanelli, F. Cardano, D. Gadler, D. Morandini, A. Piccoli, S. Salam,

A. M. Sharear, A. Ventura, P. Abrahamsson, and D. T. Oyetoyan, "Bringing the cloud to rural and remote areas via cloudlets," in *Proceedings of the 7th Annual Symposium on Computing for Development*, 14. 2016.

[65] D. Taibi, V. Lenarduzzi, C. Pahl, and A. Janes, "Microservices in agile software development: a workshop-based study into issues, advantages, and disadvantages," in *XP2017 Scientific Workshops*, 2017.

[66] N. C. Mendonca, P. Jamshidi, D. Garlan, and C. Pahl, "Developing Self-Adaptive Microservice Systems: Challenges and Directions," in *IEEE Software*. 2020.

[67] D. Taibi, V. Lenarduzzi, and C. Pahl, "Microservices Anti-Patterns: A Taxonomy," in *Microservices - Science and Engineering*, Springer. 2019.

[68] A. Samir and C. Pahl, "Anomaly Detection and Analysis for Clustered Cloud Computing Reliability," in *Intl Conf on Cloud Computing, Grids, and Virtualization*. 2019.

[69] A. Samir and C. Pahl, "A Controller Architecture for Anomaly Detection, Root Cause Analysis and Self-Adaptation for Cluster Architectures," in *Intl Conf on Adaptive and Self-Adaptive Systems and Applications*. 2019.

[70] J. Ehlers, A. van Hoorn, J. Waller, and W. Hasselbring, "Self-adaptive software system monitoring for performance anomaly localization," in *IEEE International Conference on Autonomic Computing, ICAC*, 197–200, 2011.

[71] C. Pahl, "Layered ontological modelling for web service-oriented model-driven architecture," in *European Conference on Model Driven Architecture-Foundations and Applications*. 2005.

[72] S. Murray, J. Ryan, and C. Pahl, "A tool-mediated cognitive apprenticeship approach for a computer engineering course," in *Proceedings 3rd IEEE International Conference on Advanced Technologies*, 2-6. 2003.

[73] C. Pahl, R. Barrett, and C. Kenny, "Supporting active database learning and training through interactive multimedia," in *ACM SIGCSE Bulletin* 36 (3), 27-31. 2004.

[74] C. Kenny and C. Pahl, "Automated tutoring for a database skills training environment," in *ACM SIGCSE Symposium 2005*, 58-64. 2003.

[75] X. Lei, C. Pahl, and D. Donnellan, "An evaluation technique for content interaction in web-based teaching and learning environments," in *IEEE International Conference on Advanced Technologies*, 294-295. 2003.

[76] M. Melia and C. Pahl, "Constraint-based validation of adaptive e-learning courseware," in *IEEE Transactions on Learning Technologies* 2(1), 37-49. 2009.

[77] R. Nathuji, A. Kansal, and A. Ghaffarkhah, "Q-clouds: Managing performance interference effects for QoS-aware clouds," in *European Conference on Computer Systems*, 237–250. 2010.

[78] F. Ghirardini, A. Samir, I. Fronza, and C. Pahl, "Model-Driven Simulation for Performance Engineering of Kubernetes-style Cloud Cluster Architectures," in *ES-OCC 2018 Workshops, PhD Symposium, EU-Projects*, 2019.

Trajectory Regulation for Walking Multipod Robots

Jörg Roth
Nuremberg Institute of Technology
Faculty of Computer Science
Nuremberg, Germany
e-mail: Joerg.Roth@th-nuernberg.de

Abstract— **Walking on a computed path is a fundamental task for mobile robots. Due to error effects such as slippage or imprecise mechanics, motion commands usually cannot exactly be executed. Thus, resulting positions and orientations may differ from the expectations. A robot has to compensate motion errors and should create a continuous movement that minimizes the difference between planned and real positions. This problem becomes even more difficult in case we have legged mobile robots instead of wheeled robots. In this paper, we present different mechanisms to regulate walking for multipod robots such as hexapods. They are based on virtual odometry, slippage detection and compensation as well as different types of regulation trajectories. These mechanisms are implemented and tested on the Bugbot hexapod robot.**

Keywords – Multipods; Hexapod; Autonomous Walking; Path-Following; Trajectory Regulation.

I. INTRODUCTION

A significant task of mobile robots is to navigate and move to a target position. The navigation in partly unknown environments is well-understood – we know numerous approaches to generate paths based on environment maps, created from sensor input. For the actual moving task, however, there is the problem of imprecise execution of movement commands. The problem is even more difficult, if we have walking robots such as hexapods as the movement consists of several phases with different motor actions. This is a major difference to driving robots, as their wheel-based movement typically can be fully controlled by steering angles or motor revolutions per time.

Figure 1 illustrates the problem. A hexapod should walk on a straight line. A slippery area on the ground (not known by the robot) has a lower traction because of, e.g., ice. If the robot walks without considering this area, the right feet provide smaller propulsion. As a result, the real trajectory is a curve to the right, not straight as intended.

As a desired behaviour, the robot should automatically adapt to the respective ground and execute a left curve to move back to the planned trajectory. On the first view, this problem is a typical regulation problem addressed by control theory. We could consider the joint angles (e.g., 18 for a 6-legged hexapod with 3 servos per leg) as the system output and want to minimize the difference of real and desired location and orientation. However, we have two significant differences to the classical problem.

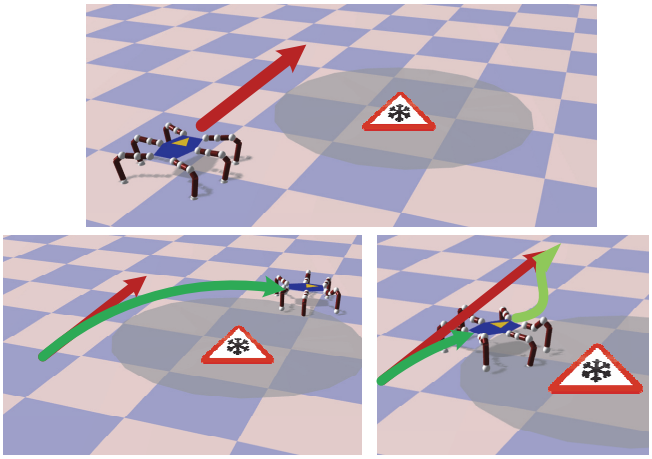


Figure 1. Illustration of the regulation problem

First, the regulation task is heavily influenced by non-holonomic walking constraints and obstacles in the environment. Second, the control output is not a set of joint angles, but a *sequence* of different joint angles, i.e., a function of angles over time (i.e., the *gait*). As a result, traditional tools to relate the joints speed and the resulting robot location (e.g., Jacobian matrices) are not suitable.

This paper presents a regulation approach based on the following ideas:

- We introduce the concept of *virtual odometry* to abstract from complex walking gaits.
- We measure and compensate *slippage* as main source of disturbance.
- We periodically compute *regulation trajectories* to a pose ahead on the formerly planned path.
- As the regulation trajectories are not computed in zero-time, we introduce *micro regulation* to compensate the delay.

This paper is an extended version of a shorter publication presented at the ADAPTIVE 2019 [1]. It extends the original paper in the following ways: First, we demonstrate the approach in more depth, in particular the mathematics behind it. Second, we introduce the concept of micro regulation to address the delay problem. Third, we provide a more detailed evaluation of the overall approach.

In Section II, we present related work. Section III describes our regulation approach. Experiments are presented in Section IV. Section V concludes the paper.

II. Related Work

Research on path following and trajectory tracking has a long tradition in control theory [2][3][4]. The basic goal is to provide a formal representation of the so-called *control law* [5]. Both, the vehicle and trajectories are strongly formalized in order to derive quality statements, in particular regarding the controller's stability [6].

Model Predictive Control (*MPC*) [7][8] is based on a finite-horizon continuous time minimization of predicted tracking errors. At each sampling time, the controller generates an optimal control sequence by solving an optimization problem. The first element of this sequence is applied to the system. The problem is solved again at the next sampling time using the updated process measurements and a shifted horizon.

Sliding Mode Control (*SMC*) is a nonlinear controller that drives system states onto a sliding surface in the state space [9][10]. Once the sliding surface is reached, sliding mode control keeps the states on the close neighbourhood of the sliding surface. Its benefits are accuracy, robustness, easy tuning and easy implementation.

The *Line-of-Sight* path following principle leads a robot towards a point ahead on the desired path. It is often used for vessels [11] or underwater vehicles [12]. The approaches differ how to reach the point ahead. Examples are arcs, straight lines or Dubins paths [13].

Another approach explicitly measures and predicts slippage, in particular of wheeled robots. As this is often a main source of disturbance to follow a path, it is reasonable to model it explicitly. In [14], effects on motors are measured for this. [15] uses GPS and inertial sensors and applies a Kalman filter to estimate slippage.

The majority of vehicles that are considered for the path following problem are wheeled robots, because their behaviour can be formalized easily. Multipods are only rarely taken into account. [16] presents trajectory planning and control for a hexapod that mainly keeps the robot balanced in rough terrain.

Pure Pursuit describes a class of algorithms that project a position ahead on the planned trajectory and create a regulation path (e.g., an arc) to reach this position. Early work about Pure Pursuit is [17]. The basic version only tries to reach a position ahead without considering the robot's orientation [18]. Improvements dynamically adapt the look-ahead distance [19].

Only rare approaches directly address the regulation problem of multipods. [20][21] consider the influence of disturbances such as grip on a climbing surface. The regulation is more fine-grained, i.e., on motor-level. As a result, the joint angles are controlled to execute a desired gait. The approach does not intend to regulate the robot movement to hold a certain trajectory or pose.

III. The Regulation Approach

Our regulation approach is embedded into the larger *Bugbot* project [22][23] (Figure 2). Bugbot is a hexapod robot, created to explore motion, navigation, world modelling and action planning for walking robots.

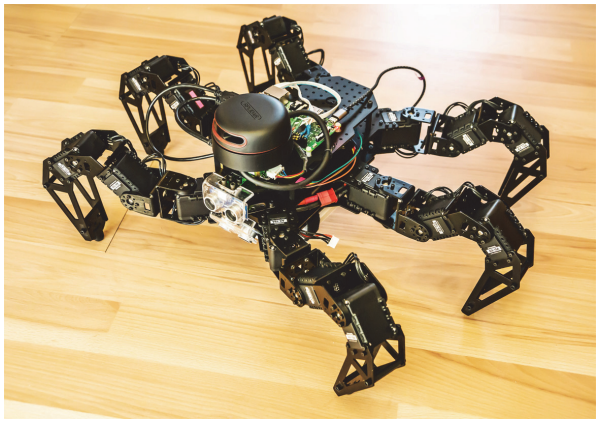


Figure 2. The Bugbot

The platform is an 18-DOF hexapod based on Trossen PhantomX Mark III. We added a Lidar device (light detection and ranging) and further sensors for collision detection and falling prevention. A Raspberry PI 3B is used for main computations, e.g., route and trajectory planning, SLAM and trajectory regulation.

Even though the Bugbot is fixed to six legs, the regulation approach is suitable for all multipods with *statically stable gaits* [24], e.g., spider-like octopods, but also robots with odd leg numbers, maybe mounted circular around the centre.

The Bugbot also comes along with a software stack (Figure 3). We have the following major components:

The *Robot Application* contains the actual task code for the robot's mission, e.g., to explore the environment, to carry things or to move to a target location.

Navigation provides a point-to-point route planning in the workspace (i.e., without dealing with the robot's orientation). This component does *not* consider non-holonomic constraints – these are shifted to lower components. It computes a line string of minimal costs that avoids obstacles. This component is useful to segment the overall path planning task into subtasks with lower complexity.

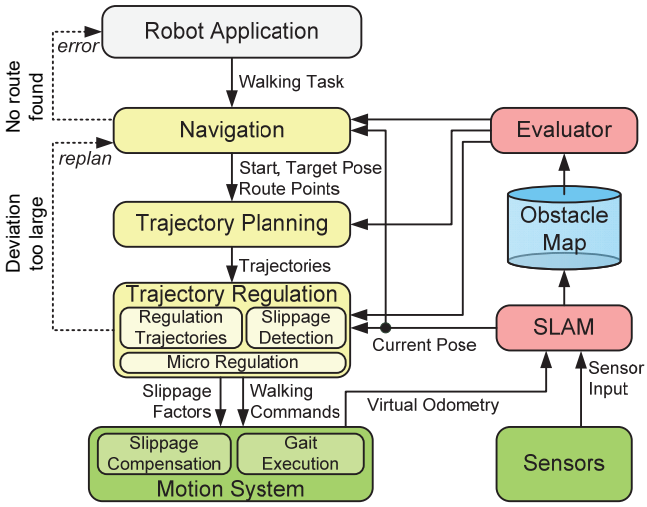


Figure 3. Data flow to execute walking tasks

Trajectory Planning computes a walkable sequence of trajectories according to the formerly computed route points and now considers non-holonomic constraints. The trajectories are taken from a set of *primitive trajectories* such as straight forward or arc. We consider primitive trajectories as directly executable by the Motion System.

Trajectory Regulation permanently tries to hold the planned trajectories, even if the position drifts off. It contains the sub-components *Regulation Trajectories*, *Slippage Detection* and *Micro Regulation* that are described in the next sections.

Simultaneous Localization and Mapping (*SLAM*) constantly observes the environment and computes the most probable own location and locations of obstacles with the help of, e.g., ultrasonic sensors or Lidar. The current error-corrected configuration is passed to all planning components. Observed and error-corrected obstacle positions are stored in an *Obstacle Map* for further planning tasks.

The *Evaluator* computes costs of routes and trajectories based on the obstacle map and the desired properties. Cost values may take into account the path length, walking time, or expected energy consumption. In addition, the distance to obstacles could be considered, if, e.g., we want the robot to keep a safety distance whenever possible.

The *Motion System* is able to execute and supervise walking commands by formalized gaits. It considers slippage and provides *Virtual Odometry*. These concepts are also described in the next sections.

In this paper, we assume that *Navigation*, *Trajectory Planning* and *SLAM* already exist. We may use an approach as described in [22] for this. We here focus on the component *Trajectory Regulation*.

When starting with the Bugbot project in 2017, we already had long-term experience with a former robot, the wheeled *Carbot* [25]. Carbot also provided navigation, trajectory planning and motion components. However, the legged Bugbot was not as easy to formalize as a wheeled robot. Moreover, we had to face issues that made it difficult to apply traditional approaches based on control theory. Walking is in general more error-prone than driving. As a result, we cannot execute regulation trajectories as precisely as expected. As different multipods may be different in the capabilities to execute certain gaits, our goal was to consider the set of possible walking commands as black box. In particular, we did not want to restrict our general regulation mechanism by specific kinematic properties, e.g., by certain gaits or leg configurations.

Our approach should directly consider obstacles and arbitrary cost functions, again given as black boxes. A certain regulation trajectory may not only be based on regulation parameters, but also on the environment.

The resulting approach was inspired by the pure pursuit idea. We project the current position ahead to the target and try to get there. But we extended the basic idea in two ways:

- We try to reach a planned *configuration*, i.e., position *and* orientation. This is much more difficult than only to reach a planned position, but as a benefit, future configurations are much closer to the intended path.
- We are not restricted to a certain primitive trajectory (e.g., single arc) to reach the configuration ahead, but execute a full trajectory planning step that may result in multiple primitive trajectories.

We use the trajectory planning both to compute a full path to the final target, as well as for the regulation approach. As a benefit, both components produce output that the Motion System directly accepts as walking commands. Moreover, the motion capabilities are modeled in one place in the system. However, we have to face the following issues:

- As we do not explicitly model a control law, we have to consider sources of disturbance, foremost the slippage effect.
- As the regulation component permanently calls a trajectory planning, we have to consider execution time. In our approach, we thus apply an efficient trajectory planning approach.
- Even though executed fast, the trajectory planning is not performed in zero time. Thus, the robot slightly has moved, before the next trajectory is computed. We need a further mechanism, we call *micro regulation*, to compensate this effect.

In Figure 3, we also see a facility to report errors to the components above. If planned and real pose deviate too much, we consider the regulation as failed and restart navigation. With a newly planned route, the first deviations are small. With this facility, actually even naïve implementations of the Trajectory Regulation would work in theory, but may have bad performance in reality. In particular, we then had to face the *cascading failure problem* as described in Section III.E.

Before we describe the regulation approach in detail, we start with the motion model and mathematical foundations.

A. The Motion Model

Multipod robots often have legs as shown in Figure 4. Legs must have at last 3 degrees of freedom to freely place and move the foot during gait execution. The leg segments usually are called *Coxa*, *Femur* and *Tibia* based on insect anatomy naming. Robot legs with more degrees may provide redundancy in leg positioning, but are not generally capable to execute more gaits. In this paper, we abstract from inverse kinematics questions and assume the controlling mechanism is capable to place the feet as required by a movement.

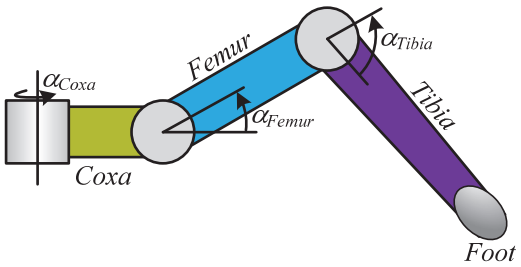


Figure 4. Typical construction of a multipod leg

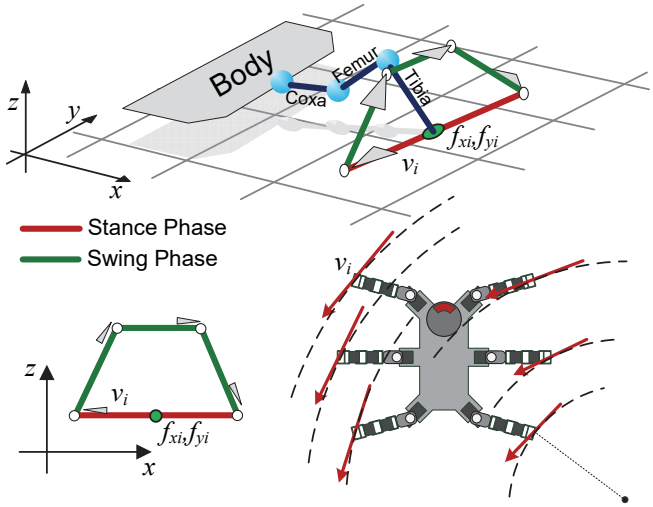


Figure 5. Structure of a multipod gait (top and left), v_i and arc trajectory (bottom right)

Multipods can walk in different ways. First, we can look at the actual trajectory, e.g., straight forward, sideways (i.e., crab gait), arc or turn in place. Second, we can distinguish the *gait* that defines the time sequence of moving legs.

We start with the trajectory. Each leg moves through two phases when walking:

- In the *stance phase,* the foot touches the ground. This phase is important for the robot's static stability. Always at least three feet must be in the stance phase, whereas the feet's convex hull must enclose the point below the centre of gravity.
- In the *swing phase,* the foot is lifted and moved to the start of the stance phase. During the swing phase, the leg can move over small obstacles.

Figure 5 shows the two phases for a specific leg. Let (f_{xi},f_{yi}) denote the neutral foot position of leg i. It marks the centre of a stance movement vector v_i in local robot coordinates. In world coordinates, the foot remains on the ground at the same position (in the absence of slippage). We assume the stance movement is linear or can at least be linearly approximated.

In the swing phase, the leg is lifted and moved in walking direction, whereas the ground projection may reside on v_i. The swing phase has a polygonal representation in three dimensions, but can be defined by a 2D polygon roto-translated to be aligned to v_i.

The second facet of multipod walking is the *gait*. Gaits define the cooperation of legs in the respective phases. Figure 6 shows the example of the *Ripple* gait, a gait that always has two lifted legs. Many further gaits are known, e.g., *Tripod* or *Wave* that differ in stability and propulsion [24]. A gait is fully described by a *gait matrix* that specifies per leg (row) and step (column) whether a leg is in swing phase (1) or stance phase (0). For the Ripple gait, we get the matrix

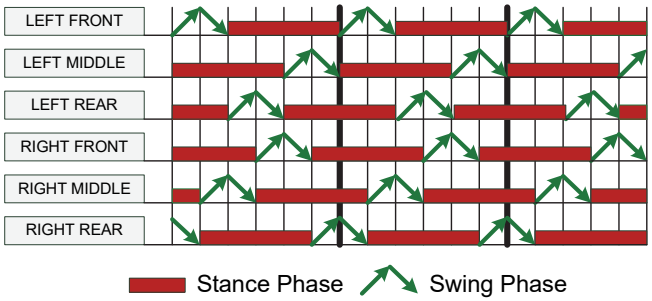


Figure 6. Gait pattern for the Ripple gait of hexapods

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (1)$$

We assume that gait execution and the choice for a certain gait are encapsulated in the Motion System component. For this, it receives the gait matrix and stance vectors v_i and is able to autonomously execute a gait as long as required by the respective trajectory.

An important observation: we can deal with movement trajectory and gait independently. This means that the respective trajectory shape is *not* influenced by the gait sequence pattern. The gait only affects the movement speed and stability.

B. Mathematical Foundations

For the intended approach, we need mathematical answers for three questions:

(A) Given a primitive trajectory, what are the stance vectors v_i to move along the required trajectory?

(B) Given two robot poses, what is the primitive trajectory that connects these poses?

(C) Given stance vectors v_i, what is the primitive trajectory the robot walks?

(A) and (C) are reverse questions. (B) and (C) are similar, but base on different input variables. We additionally could ask for the pose after walking a primitive trajectory. This would be the reverse of (B). However, this usually has a simple solution.

In the following, we consider two primitive trajectories: moving straight to target (t_x, t_y) and moving along an arc with centre c_x, c_y and curve radius r. More trajectories are conceivable, e.g., moving along clothoids. However, these have certain properties that are more suitable for driving robots. We also consider turning in place as primitive, but subsume it under the arc trajectory (whereas the arc centre is the robot centre).

We further assume that there exists a maximum stance vector length v_{max}. The maximum length is a result of the respective leg mechanics, e.g., leg segment lengths, servo angle limits and collision areas between the legs.

The solution (A) for straight trajectories is

$$v_i = -\begin{pmatrix} t_x \\ t_y \end{pmatrix} \frac{v_{\max}}{\left\| \begin{pmatrix} t_x \\ t_y \end{pmatrix} \right\|} \tag{2}$$

For arcs, we have the following constraints. First, the stance vector must be orthogonal to the line between neutral position and arc centre. Second, the ratio of stance vector lengths of two legs must be equal to the ratio of the distances between the respective neutral positions and arc centre. Third: the largest stance vector must have length v_{max}. More formally:

$$v_i \times \begin{pmatrix} f_{ix} - c_x \\ f_{iy} - c_y \end{pmatrix} = 0 \,, \frac{\|v_i\|}{\|v_j\|} = \frac{\left\| \begin{pmatrix} f_{ix} - c_x \\ f_{iy} - c_y \end{pmatrix} \right\|}{\left\| \begin{pmatrix} f_{jx} - c_x \\ f_{jy} - c_y \end{pmatrix} \right\|} \tag{3}$$
$$\max(\|v_i\|) = v_{\max}$$

With these equations, we can easily construct the v_i: we turn the line between neutral position and arc centre by 90°. Then we identify leg i that has the largest distance to the arc centre – this must receive the stance vector length v_{max}. Finally, we rescale the stance vector lengths according to the required relation, i.e.,

$$v_i' = \begin{pmatrix} c_y - f_{iy} \\ f_{ix} - c_x \end{pmatrix}, v_i = v_i' \cdot \frac{v_{\max}}{\max(v_i')} \tag{4}$$

For (B) we want to connect the current pose with a target pose (t_x, t_y, t_θ). For t_θ=0 we get a straight trajectory to (t_x, t_y). For $t_\theta \neq 0$ we get an arc with angle t_θ. Figure 7 shows the details.

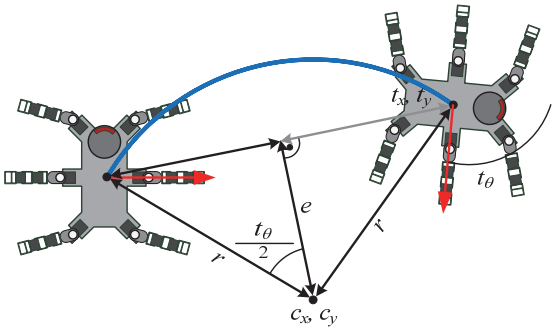


Figure 7. Construction of the arc geometry from two poses

We get the arc centre, if we add vectors $(t_x, t_y)/2$ and e, whereas e has the length

$$\|e\| = \left\| \begin{pmatrix} t_x \\ t_y \end{pmatrix} \right\| \frac{1}{2 \cdot \tan(t_\theta / 2)} \tag{5}$$

for f=1/(2tan(t_α/2)) we thus get

$$\begin{pmatrix} c_x \\ c_y \end{pmatrix} = \begin{pmatrix} t_x / 2 \\ t_y / 2 \end{pmatrix} + \begin{pmatrix} -t_y \cdot f \\ t_x \cdot f \end{pmatrix} \tag{6}$$

or

$$\begin{pmatrix} c_x \\ c_y \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1/\tan(t_\theta/2) \\ 1/\tan(t_\theta/2) & 1 \end{pmatrix} \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{7}$$

From (c_x, c_y), we finally get r.

Problem (C) is the hardest to solve. We first compute the amount of propulsion in a small time interval Δt. Even though a foot in stance phase remains on the ground, in local robot coordinates it moves along v_i. In turn, the neutral foot position (f_{ix}, f_{iy}) moves along $-v_i$ in world coordinates. When a foot remains in the stance phase during the time t_{st}, the foot moves during Δt

$$-\frac{\Delta t}{t_{st}} v_i \tag{8}$$

The time t_{st} depends on the respective gait (e.g., Tripod or Wave) and can be derived from the gait matrix: it is the ratio of zeros in a row, multiplied by total cycle time. Note: even though a foot in swing phase provides no actual propulsion to the robot, the respective (f_{ix}, f_{iy}) virtually move continuously, also for $t > t_{st}$. As a result, we have a constant speed over all steps.

Viewed from the first position, the feet virtually move from $f_i=(f_{ix}, f_{iy})$ to $f'_i=(f'_{ix}, f'_{iy})$ whereas

$$\begin{pmatrix} f'_{ix} \\ f'_{iy} \end{pmatrix} = \begin{pmatrix} f_{ix} \\ f_{iy} \end{pmatrix} - \frac{v_i}{t_{st}} \tag{9}$$

To find the respective primitive trajectory, we need a function Ψ that computes

$$\begin{pmatrix} t_x \\ t_y \\ \alpha \end{pmatrix} = \Psi\left((f_1 \ldots f_n), (f'_1 \ldots f'_n) \right) \tag{10}$$

Ψ denotes a function to compute a roto-translation, which maps all positions of the first list to positions of a second list, meanwhile minimizing the mean square error. We apply an approach based on Gibbs vectors [26] for Ψ: We look for a rotation matrix R and translation t with

$$f'_i = R \cdot f_i + t \tag{11}$$

According to the Cayley transform [27], we are able to replace the rotation matrix as follows

$$f_i' = (I+Q)^{-1}(I-Q)\cdot f_i + t \tag{12}$$

where I is the unity matrix and $Q=[q]_\times$ with q the vector of Rodriguez parameters (Gibbs vector), where $[]_\times$ denotes the cross product operation by a matrix, i.e., $a\times b=[a]_\times \cdot b$.

We can rewrite this to

$$\begin{aligned} f_i'(I+Q) &= (I-Q)\cdot f_i + t(I+Q) \\ &= (I-Q)\cdot f_i + t^* \end{aligned} \tag{13}$$

where $t^*=t(I+Q)$. We get

$$f_i' - f_i = -Q(f_i + f)_i + t^* \tag{14}$$

For two dimensions we get

$$\begin{aligned} y = \begin{pmatrix} f_{1x}' - f_{1x} \\ y_{1y} - f_{1y} \\ \dots \\ f_{nx}'' - f_{nx} \\ f_{ny}' - f_{ny} \end{pmatrix} &= \begin{pmatrix} f_{1y} + f_{1y}' & 1 & 0 \\ -(f_{nx} + f_{nx}') & 0 & 1 \\ & \dots & \\ f_{1y} + f_{1y}' & 1 & 0 \\ -(f_{nx} + f_{nx}') & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} g_\alpha \\ t_x^* \\ t_y^* \end{pmatrix} \\ &= H \cdot \begin{pmatrix} g_\alpha \\ t_x^* \\ t_y^* \end{pmatrix} \end{aligned} \tag{15}$$

This is an overdetermined linear equation system. We minimize least squares with

$$\begin{pmatrix} g_\alpha \\ t_x^* \\ t_y^* \end{pmatrix} = (H^T H)^{-1} H^T y \tag{16}$$

We finally get the roto-translation (t_x, t_y, α) with

$$\begin{aligned} \alpha &= 2\cdot \arctan(g_\alpha) \\ \begin{pmatrix} t_x \\ t_y \end{pmatrix} &= \begin{pmatrix} \frac{1}{1+\tau^2} & \frac{\tau}{1+\tau^2} \\ \frac{-\tau}{1+\tau^2} & \frac{1}{1+\tau^2} \end{pmatrix} \cdot \begin{pmatrix} t_x^* \\ t_y^* \end{pmatrix} \end{aligned} \tag{17}$$

where $\tau = \tan(\alpha/2)$.

C. Computing Trajectories

A basic building block for the trajectory regulation is to compute regulation trajectories, i.e., such trajectories that bring the robot back to a planned path. As a basic idea, we use for regulation trajectories and long-range paths the same approach [22]. We adapted the original approach that was optimized for a wheeled robot [25] to a walking robot. It is based on the following ideas:

- The navigation component (Figure 3) solely operates on workspace $\mathcal{W}$ and computes a sequence of collision-free lines of sight (with respect to the robot's width) that minimize the costs.
- As the navigation only computes route points in $\mathcal{W}$, we have to specify additional variables in $\mathcal{C}$ (here orientation θ). From the infinite assignments, we only consider a small finite set.
- From the infinite set of trajectories between two route points, we only consider a finite set of *maneuvers*. Maneuvers are sequences of primitive trajectories, for which we know formulas that derive the respective parameters (e.g., curve radii) from start and target configurations.
- Even though these concepts reduce the problem space to a finite set of variations, this set would by far be too large for complete checks. We thus apply a Viterbi-like approach that significantly reduces the number of checked variations to find an optimum.

We carefully separated the cost function (component *Evaluator*, Figure 3) from planning components. We assume that there is a mapping from a route or trajectory sequence to a cost value according to two rules: first, we have to assign a single, scalar value to a trajectory sequence that indicates its costs. If costs cover multiple attributes (e.g., walking time *and* battery consumption), the cost function has to weight these attributes and create a single cost value. Second, a collision with obstacles has to result in infinite costs.

The basic capabilities of movement are defined by the supported set of primitive trajectories. The respective set can vary between different robots. A walking robot should support:

- $L(\ell)$: linear (straight) walking over a distance of ℓ;
- $T(\Delta\theta)$: turn in place over $\Delta\theta$;
- $A(\ell, r)$: move a circular arc with radius r (sign distinguishes left/right) over a distance of ℓ

We are able to map primitive trajectories directly to walking commands that are natively executed by the robot's Motion System (Figure 3).

A certain multipod may also support *holonomic* locomotion, e.g., walk straight and turn the viewing direction during a single motion command. In this case, however, we have changing stance vectors over time, what complicates the gait formalism. Moreover, for a certain walking scenario not all trajectories may be reasonable. E.g., we may expect a camera, an ultrasonic sensor or the sensors that prevent from falling downstairs to always point in walking direction. Thus, we require identical viewing and walking directions for all trajectories of type *A* or *L*, i.e., the robot only moves in the direction, it also views. Considering these constraints, it is not possible to reach a certain pose with a single primitive trajectory. At this point, we introduce *maneuvers*. Maneuvers are small sequences of primitive trajectories (usually 2-3 elements) that are able to map given start and target configurations c_s, $c_t \in \mathcal{C}$. More specifically:

- A maneuver is defined by a sequence of primitive trajectories (e.g., denoted *ALA* or *AA*) and further

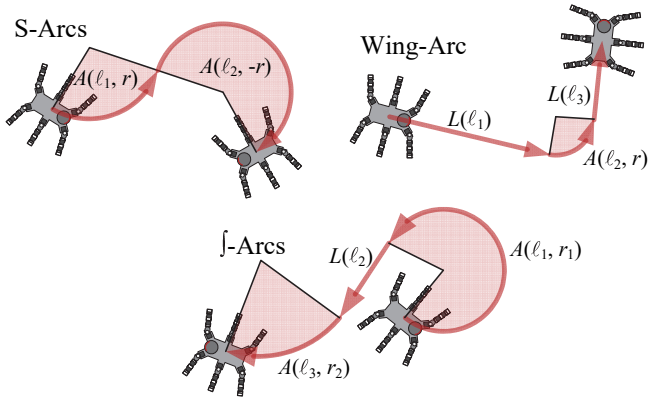


Figure 8. Example maneuvers

constraints. Constraints may relate or restrict the respective primitive trajectory parameters.

- For given c_s, $c_t \in \mathcal{C}$ there exist formulas that specify the parameters of the involved primitive trajectories, e.g., ℓ for L, A and ℓ, r for A.
- Sometimes, the respective equations are underdetermined. As a result, multiple maneuvers of a certain type (sometimes an infinite number) map c_s to c_t. Thus, we need further parameters, we call *free parameters* to get a unique maneuver.

Figure 8 shows three example maneuvers: S-Arcs (*AA*), Wing-Arc (*LAL*) and ∫-Arcs (*ALA*). For, e.g., S-Arcs we have to set up equations to get the respective free parameters ℓ_1, ℓ_2, r. To simplify the computation, we first roto-translate start and target to move the start to (0, 0, 0) and target to (x_t', y_t', θ_t'). We get

$$r = \frac{y_t'(1+ct) - x_t' st - \sqrt{x_t'^2 (st^2 - 2ct + 2) + y_t'^2 (3 + ct^2) - 2x_t' y_t' st(1+ct)}}{2(ct-1)}$$

$$\ell_1 = r \cdot \arccos\left(\frac{(1+ct)}{2} - \frac{y_t'}{2r}\right) \qquad (18)$$

$$\ell_2 = \ell_1 - r \cdot \theta_t'$$

where st=sin(θ_t'), ct=cos(θ_t'). We identified a total of 8 maneuvers so far (Table I). We assigned names that illustrate the maneuver's shape, e.g., the J-Bow goes through a path that looks like the letter 'J'. The Dubins-Arcs correspond to the combination with three arcs of Dubins original approach [13]. From all maneuvers, ∫-Arcs can be considered as a 'Swiss knife': it allows reaching any target configuration without a turn in place whereas the middle linear trajectory spans a reasonable distance to the target.

We also may invent new maneuvers to increase the overall walking capabilities. For a new maneuver, we only have to set up equations that derive the respective trajectory parameters from start and target configuration.

To find a trajectory sequence of maneuvers is an optimization problem. For a given start and target pose in $\mathcal{C}$ and a list of route points in $\mathcal{W}$ we have to find a sequence of maneuvers (and thus primitive trajectories) that

TABLE I. AVAILABLE MANEUVER TYPES

Maneuver	Pattern	Free Parameters
1-Turn	*LTL*	no
2-Turns	*TLT*	no
J-Bow	*LA*	arc radius
J-Bow2	*AL*	arc radius
∫-Arcs	*ALA*	two arc radii
S-Arcs	*AA*	no
Wing-Arc	*LAL*	arc radius
Dubins-Arcs	*AAA*	(same) arc radius for all three arcs

- connect start pose, route points and target pose,
- minimizes the costs, computed by the Evaluator.

The controllable variables are: the maneuver types, their free parameters and the orientation angles at the route points. From the infinite set of the respective variations, we choose a finite promising set of candidates. Even though finite, the number of variations still is by far too large for a complete check. To give an impression: for 5 route points we get a total number of 20 million, for 20 route points $2 \cdot 10^{37}$ permutations. Obviously, we need an approach that computes an appropriate result without iterating through all permutations.

Our approach is inspired by the Viterbi algorithm that tries to find the most likely path through hidden states. To make use of this approach, we replace 'most likely' by 'least costs', and 'hidden states' by 'unknown parameters'. We thus look for a sequence of maneuvers/orientations/free parameters that connect them with minimal costs. Details of the underlying algorithm can be found in [22].

This approach is suitable, because optimal paths have a property: the interference between two primitive trajectories in that path depends on their distance. If they are close, a change of one usually also causes a change of the other, in particular, if they are connected. If they are far, we may change one trajectory of the sequence, without affecting the other. Viterbi reflects these characteristics, as it checks all combinations of neighbouring (i.e., close) maneuvers to get the optimum. We can reduce the number of variations to check for a complete route to some thousands.

As a further benefit of the approach: The first primitive trajectory of the final path converges very fast. If fixed, the robot can start walking, while further trajectories are computed during the movement. This property makes our trajectory planning as an ideal candidate for regulation trajectories: The regulation trajectory is frequently computed in the background, each of it only, until the first primitive trajectory is fixed. During walking, the next primitive trajectory can be computed by following iterations.

D. Virtual Odometry

Leg movement with a complex timing pattern is difficult to handle in the context of trajectory planning and regulation. For geometric computations the model of turning wheels is more convenient, because we have a simple relation between motor revolutions and odometry. This leads to the idea of *virtual odometry*: We transform walking to corresponding wheeled movement. We could think of roller skates attached

to the multipod's legs while the legs remain in neutral position (f_{xi}, f_{yi}).

To compute virtual odometry, we intercept the formal gait description that is passed to the Motion System, namely:

- the neutral position (f_{xi}, f_{yi}) for each leg i,
- the stance vector v_i for each leg i,
- the time t_{st} that the gait resides in stance phase for a complete cycle of one stance and one swing phase.

According to formulas (10)-(17) we are able to compute a roto-translation (t_x, t_y, α) that maps the neutral leg positions to the moved neutral positions meanwhile minimizing mean squares. We further get (c_x, c_y) in case of an arc from formula (7).

We now assume in a small time interval the robot either only moves an arc (respectively turn in place) or linear trajectory. For small intervals and thus small running lengths, this is a reasonable approximation. We further consider the angular velocity and absolute speed as constant in a small time interval. If we consider both arc and straight movement, we get the moving distance for a leg i over time Δt as

$$\ell_i(\Delta t) = \Delta t \cdot \begin{cases} |\alpha| \sqrt{(f_{xi} - c_x)^2 + (f_{yi} - c_y)^2} & \text{if } \alpha \neq 0 \\ \sqrt{t_x{}^2 + t_y{}^2} & \text{if } \alpha = 0 \end{cases} \quad (19)$$

We call $\ell_i(\Delta t)$ the *virtual odometry*. It represents the *expected* portion of the overall moving distance of each foot when walking.

E. Slippage Detection and Compensation

With the help of virtual odometry we now are able to compute the expected run length and expected location. If they deviate from the planned trajectory, we try to walk back, using regulation trajectories (see below).

If we did not take into account slippage, the regulation trajectories will not be executed as expected, thus the deviation may increase (Figure 9). Even though the robot permanently tries to go back on the planned path, the distance gets larger, because the regulation trajectories were also executed with a drift (here, to the left). In this case, we get what we call the *cascading failure problem*: at a certain point, the regulation fails and triggers a new route planning by the navigation component. This however, can only fix the problem for a certain time, as also further routes are not executed as expected. Finally, the robot, e.g., moves too close to a wall and the entire movement is stopped with an error.

In addition to the expected walking distances, we now need the real distances. We assume that the robot owns sensors (e.g., Lidar) and respective SLAM mechanisms that permanently estimate the robot's real position. We consider these mechanisms as black box, but expect, they detect the real pose change (t_x', t_y', α') after walking a time Δt. We again assume that during Δt, only a single movement pattern is executed. This obviously is wrong, if there is a change in the trajectory (e.g., changing from arc to straight). However, for small Δt, we can model both patterns by a single ('average') pattern, thus receive only small errors.

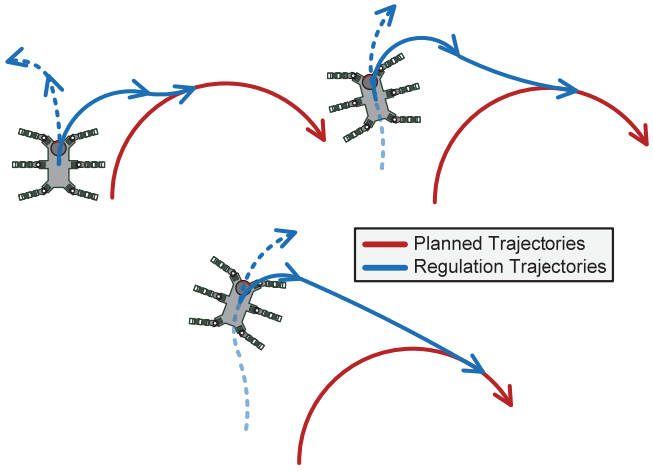


Figure 9. Problem of missing slippage compensation

For given (t_x', t_y', α') we can apply formulas (7) and (19) to get the *real* walking distances $\ell_i'(\Delta t)$. We define

$$S_i = \frac{\ell_i(\Delta t)}{\ell_i'(\Delta t)} \quad (20)$$

as the *leg-specific slippage factor* and

$$S = \frac{1}{L} \sum_i S_i \quad (21)$$

as the *general slippage factor*, where L is the number of legs. Obviously $S \geq 1$ in reality. S describes the slippage property of the current bottom's pavement. E.g., $S=2$ means that the robot walks half as far as expected when executing a certain trajectory. The S_i describe slippage *per leg* and could indicate malfunctioning leg servos or feet that do not properly touch the ground.

We are able to compensate slippage in two ways:

- only compensate the general slippage;
- also compensate leg-specific slippage.

The assumption is: what we measured recently is a good estimation for the nearer future. If, e.g., we walk on a slippery floor, we can consider the respective slippage factor when executing next trajectories, because it is likely to reside on the same floor for a certain time. To consider the general slippage factor, we have to extend the respective trajectory by the discovered factor:

- Walking straight over a certain distance, we have to multiply the planned distance by S.
- Walking on an arc, we have to multiply the planned arc angle by S.

To consider the leg-specific slippage is more difficult. The problem: these factors do not only affect the trajectory length, but also its shape. E.g., if we want to walk straight

with different factors S_i for left and right legs, the robot effectively walks on an arc instead. A first approach would be to extend the respective stance vectors. E.g., if we got S_i=2 for a specific leg (i.e., the leg produces only half of the expected propulsion), we could enlarge v_i by 2 to compensate this effect. However, this is not always possible, because the stance vector lengths are limited to v_{max} – either by the mechanics, or because neighbour legs should not collide during walking. Thus, we usually are only able to shorten the stance vectors. Our approach is to compute

$$S_{\max} = \max(S_i)\,,\ \widetilde{S}_i = S_i / S_{\max}\,,\ \widetilde{v}_i = v_i \cdot \widetilde{S}_i \qquad (22)$$

We use S_{max} as the general factor to extend the trajectory and multiply each leg's stance vector by $\widetilde{S}_i$. Note that $\widetilde{S}_i \leq 1$, thus a stance vector only can get smaller.

It depends on the respective scenario, whether the compensation only should consider the general slippage or should also apply a leg-specific compensation. The latter is only reasonable, if we expect a leg-specific slippage that may be result of malfunctioned legs or different pavement for different legs.

F. Regulation Trajectories

The task to compensate the drift during walking and to meet the planned trajectories is related to control theory, where a system tries to produce a desired output with the help of controllable input values. In the case of trajectory regulation, however, the desired output is a pose that usually cannot directly be achieved by adapting current joint angles or by a primitive walking operation. Due to non-holonomic constraints, we usually require a *sequence* of trajectories, i.e., our maneuvers.

To explain our approach, we need some definitions. First, we need a function *TP* that provides a trajectory planning from start pose *s* to target pose *t* based on Section III.C.

$$(T_i) = TP\big((s_x, s_y, s_\theta), (t_x, t_y, t_\theta)\big) \qquad (23)$$

The (T_i) is a sequence of primitive trajectories. We further need to identify an *expected* pose *e* of a current pose *c*.

$$(e_x, e_y, e_\theta) = E\big((T_i), (c_x, c_y, c_\theta)\big) \qquad (24)$$

Expected means: the intended pose on the planned path for a given pose. If the multipod remains on the planned path, *c* and *e* are identical. If the pose leaves the planned path, we have to introduce a notion of '*nearest pose on the trajectory*', whereas we may have different definitions for this. The function *E* may be *stateful* or *stateless*. A stateful implementation observes the current walking task and identifies the expected pose based on walking time or virtual odometry. As an example: we could measure the walking distance from the start of walking on (T_i) and identify the pose that has the same distance from the start. A stateless implementation only identifies the nearest trajectory point based on geometric distances. In our experiments (Section IV), we implemented the stateless version.

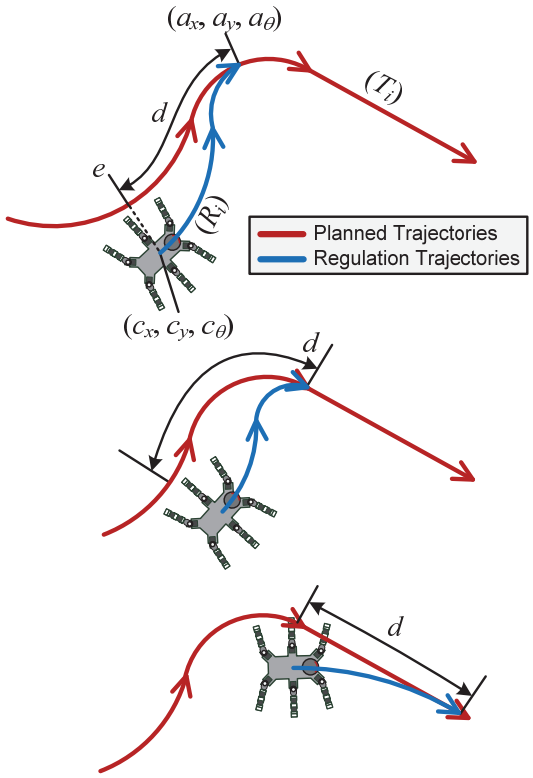


Figure 10. Idea of regulation-ahead

We finally need a function *A* that projects the current expected pose *ahead*.

$$(a_x, a_y, a_\theta) = A\big((T_i), e, d\big) \qquad (25)$$

Here, *d* describes how much the current expected pose is projected ahead in target direction. We assume *d* to be constant. Figure 10 illustrates the idea.

We now compute a trajectory sequence (R_i) that brings the robot back to the originally planned trajectory. Our approach is to compute

$$(R_i) = TP\big((c_x, c_y, c_\theta), A\big((T_i), E\big((T_i), (c_x, c_y, c_\theta)\big), d\big)\big) \qquad (26)$$

The major benefit: we do not have to introduce a new mechanism to plan regulation trajectories, but re-use the function *TP*. One could suggest to bypass regulation trajectories and directly compute *TP*(*c*, *t*). However, the pose ahead is much closer to the current pose, thus a planning is much more efficient. Furthermore, we do not expect obstacles between current and ahead pose, as the original path already is planned to be obstacle-free.

We finally have to think about *d*:

- For a small *d*, we force the robot to walk on sharp turns to restore the planned trajectory sequence.

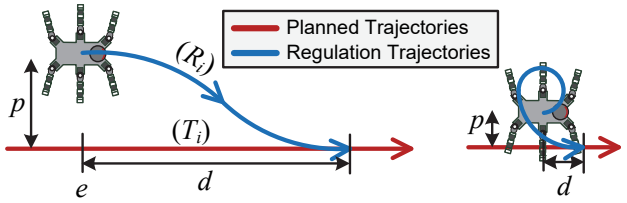


Figure 11. Effects of large and small d, p

- For a large d, the robot walks a long time parallel to the planned trajectory before it reaches the meeting point.

Both effects lead to higher costs – either because the path gets significantly longer or because the robot walks on positions with higher costs, only nearby the planned trajectory.

Figure 11 illustrates these effects. In this example, we planned a linear trajectory and the real position is besides the linear trajectory with distance p, but with correct orientation angle.

If both p and d are large, usual regulation trajectories contain two arcs. If p and d are small (Figure 11 right), the regulation trajectory may be a spiral that starts in opposite direction. This situation is unwanted, as the regulation first enlarges the distance to the planned trajectory.

We want to investigate this effect. As a first observation, it heavily depends on the walking capabilities, in particular the set of primitive trajectories and minimal arc radii, in addition the cost function. We thus cannot give a general specification of a 'good' d. However, we can provide an idea to discover d for a respective scenario.

Let $|(R_i)|$ be the length of the regulation trajectories. We define

$$q = \frac{|(R_i)|}{d} \tag{27}$$

as the *stretch factor*. It specifies how much longer the regulation path is compared to the way on the planned path. Figure 12 shows typical curves of q.

Due to the effect presented in Figure 11 (right), small d result in high q. At a certain point (here at d=40 cm) q is close to 1.0. For d>40 cm, we get only minor improvements

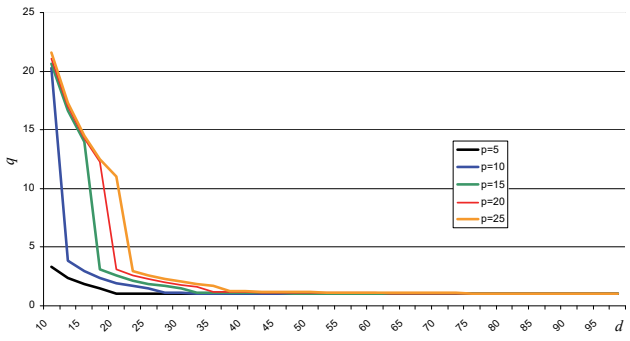


Figure 12. Typical stretch factors (d, p in cm)

of q. As a result, d=40 cm is a good choice for our scenario.

This is only an example for a certain scenario. If we want to discover an appropriate d for other scenarios, we have to consider the range of expected position errors (here p), but also the expected orientation errors.

G. Micro Regulation

We can compute (R_i) with the function *TP* periodically, e.g., every few seconds without considerable stressing the CPU. However, on very small mobile platforms that also execute additional tasks (e.g., image processing), the corresponding planning process may be delayed. This may cause a problem: *TP* is executed for a specific pose, but when *TP* finished, the robot has slightly moved to another pose. If the computed regulation trajectory then is applied to the new position, the endpoint does not reside on the originally planned trajectory. This in particular is a problem, if the orientation angle differs from the expectation.

It is not reasonable to stop the movement during *TP* computation as this would seriously disturb the continuous movement of legs. This forms the idea of another type of regulation – the *micro regulation*: We compute a short path that moves back to the regulation trajectory with a *single primitive trajectory*. Usually, it is not possible to reach a position *and* orientation with a single trajectory without to relax some constraints.

The idea of maneuvers (i.e., *multiple* primitive trajectories) respects the requirement to always walk in forward-direction and considers the non-holonomic constraint not to move side-ways. However, a walking robot *is* able to move sideways (like a crab). The idea of micro regulation is to create arc trajectories that cautiously make use of this possibility. To respect sensors that only scan in forward-direction, the amount of side-ways movement should be very small compared to forward movement.

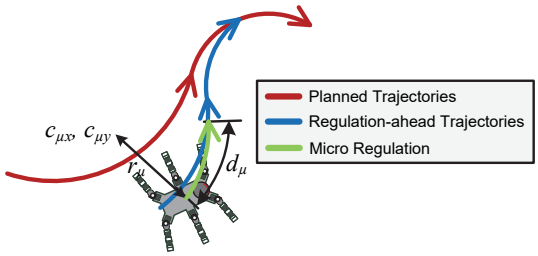


Figure 13. Idea of combined micro regulation and regulation-ahead

Figure 13 shows the idea: according to the solution of problem (B) (Section III.B), we compute a single trajectory (usually an arc) that brings the robot back to the regulation trajectory. This arc is not like the *A* trajectory (Section III.C) as the corresponding arc centre does not necessarily resides ±90° to the viewing angle. As a result, the walking direction slightly goes sideways.

Micro regulation requires fewer computation power as *TP*, as only a single trajectory has to be computed. Because we only slightly leave the regulation trajectory, we additionally may ignore obstacles. As a result, we have two background loops with different cycle times: one loop that com-

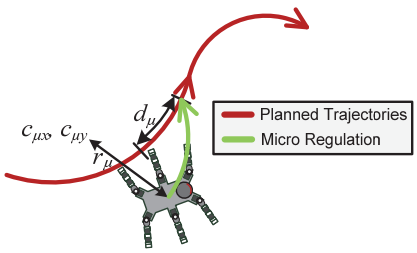


Figure 14. Idea of pure micro regulation

putes a regulation trajectory (e.g., every 4 s) and one loop that computes a micro regulation trajectory (e.g., every 2 s). In addition, the micro regulation loop has a smaller ahead distance d_μ, e.g., 20 cm. For a certain robot, the cycle time and ahead distance again must be subject to experimental optimization.

We can even go one step further (Figure 14): We do not necessarily have to compute a micro regulation trajectory to the regulation trajectory (R_i). We instead could try to reach to planned trajectory (T_i). We call the approach in Figure 14 *pure micro* regulation, in contrast to *ahead micro* (Figure 13). In case of pure micro, however, we expect a larger amount of sideways walking, as there is a greater distance between real pose and desired target pose. This effect is investigated with experiments in the next section.

IV. EXPERIMENTS

We implemented our trajectory regulation approach on the *Bugbot* platform. Even though we fully tested the approach on this platform, it was difficult to create a huge number of different experiments in reality. E.g., it is costly to test the slippage detection for different floors and different slippage factors. It is even a problem to create pavements with a very specific constant slippage factor. It is also a problem to adjust leg-specific slippage in reality in a fine-granular manner. Moreover, it is very difficult to create reproducible test runs as slippage and traction vary over time, even for the same pavement. This makes it difficult to compare results. We thus created a simulation environment that simulates the Bugbot on hardware- and physical level. A physical simulation component is able to compute gravitation, any form of slippage and collision effects. The control software is the same as on the real hardware, i.e., the simulator's Bugbot model is able to create sensor values and carries out native servo commands.

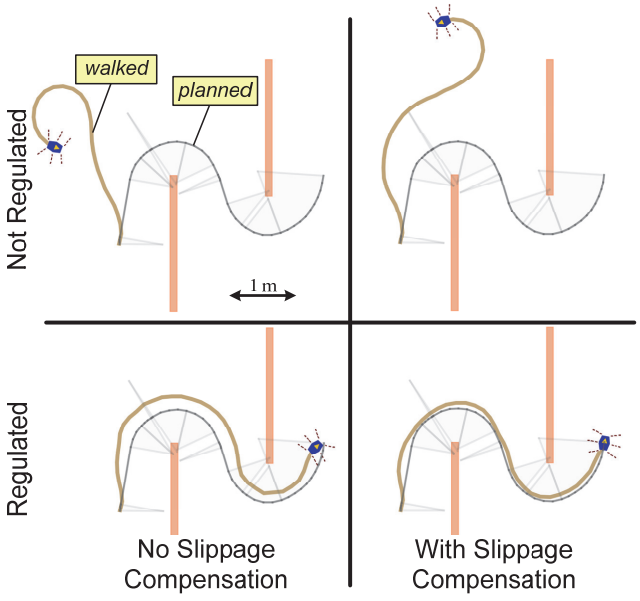


Figure 15. Simple walking scenario

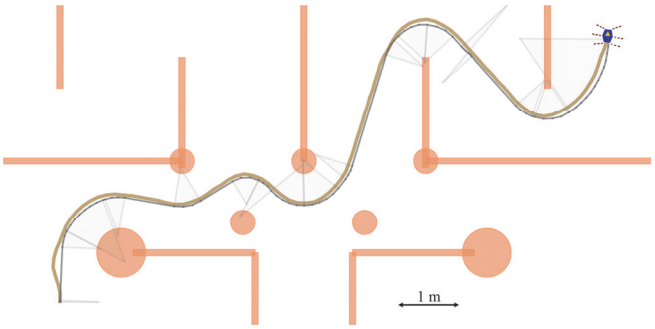


Figure 16. Complex walking scenario (regulated, slippage compensation)

We divided the experiments in two groups. The first group shows the effects of regulation vs. no regulation and slippage compensation vs. no compensation. The second group shows the effect of the different regulation approaches.

A. Regulation and Compensation

Figure 15 shows an example to illustrate the effects of slippage compensation and regulation. For regulation, we applied the regulation-ahead approach as described in Section III.C. We simulated a leg-specific slippage of 2.0 for the three left legs. This means that without any compensation, the robot walks a left arc when planned to walk right (Figure 15 top, left). With slippage detection and compensation, the shape of the planned path is mainly represented. But because the compensation is applied not before a small learning phase, the shape is rotated at the beginning (Figure 15 top, right).

Figure 15 bottom shows the regulation. On the left we see an effect when the regulation tries to meet the planned path. Because the regulation trajectories are not executed properly, we see a constant offset. On the right, we finally see both mechanisms – after a learning phase, the planned trajectory is reproduced very precisely.

Figure 16 shows a more complex example. Here, we again assigned a leg-specific slippage of 2.0 (only left legs) and in addition a general slippage of 2.0. This represents a very difficult scenario. If regulation and slippage compensation were applied, we can see a high congruence of planned and walked path.

B. Comparison of Regulation Approaches

The second group of experiments investigates the properties of *regulation-ahead*, *pure micro* and *ahead micro*. For this, we modified the complex walking scenario above and added different zones of slippage (Figure 17).

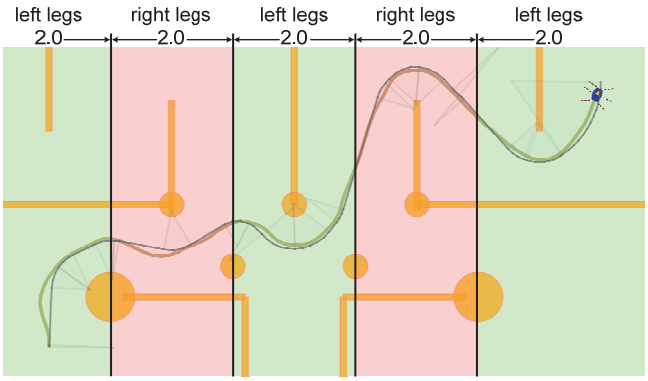


Figure 17. Walking scenario with changing slippage

We again assigned a general slippage of 2.0, but alternated the leg-specific slippage (again 2.0) every 2 m between left to right legs. This stressed the regulation mechanism: after adapting to a certain slippage, it significantly changes and the mechanism first has to learn the new situation.

We applied slippage detection and compensation and only changed to way to compute regulation trajectories. We tested the three types as shown in Table II.

TABLE II. PARAMETERS FOR THE TEST CASES

Regulation approach	d (ahead)	cycle time (ahead)	d_μ	cycle time (micro)
regulation-ahead	40 cm	4 s	n/a	n/a
micro (pure)	n/a	n/a	20 cm	2 s
micro (ahead)	40 cm	4 s	20 cm	2 s

For the ahead distance d, we selected the value from Section III.F. As we have a maximum speed of approx. 9 cm/s, a cycle time of 4 s is sufficient. For the micro regulation we choose half values, both for the ahead distance d_μ and cycle time. For the three test runs we measured three values:

- The distance d_{err} between real position and the nearest position on the planned path.
- The absolute angle error α_{err} between real orientation and the trajectory direction of the nearest planned path point.
- The absolute heading error h_{err}: it is the difference between viewing direction and walking direction. E.g., 0° means walking strictly forward, 90° means walking sideways in crab gait.

Figure 18 shows the results. We can easily see the decrease of errors d_{err} and α_{err} after adapting to the slippage. Whenever the robot enters a new area of slippage, we can see increasing errors. The errors are considerable low, even in worst case. The characteristics of h_{err} are different whether micro regulation is applied or not. If not, h_{err} is always zero. This is because TP already ensures that (R_i) only contains trajectories with forward heading. Micro regulation tries to compensate angle errors with trajectories that continuously change the heading with arcs. Not surprisingly, we thus can see a strong correlation of α_{err} and h_{err}.

Table III shows the averages of the respective values. In addition, we measured the computation time to compute regulation trajectories (R_i), micro regulation, or both in % of the overall CPU time. As we have a considerably long time between the respective computations (2 or 4 s), the total amount of time is very low.

TABLE III. TEST RESULTS

Regulation approach	CPU load in %	avg(d_{err}) in cm	avg(α_{err}) in °	avg(h_{err}) in °
regulation-ahead	0.00855	2.9	4.6	0
micro (pure)	0.00550	1.5	7.9	8.2
micro (ahead)	0.0131	1.9	4.5	3.5

Looking at the error values for all three types, the distance of real and planned position is very small. The angle errors α_{err} and h_{err} are more significant. If we have strong demands according the orientation, pure micro is not recommended as we have a maximum of 8° both for orientation and heading error. In summary, ahead micro provides the best results, but slightly requires more CPU load than regulation-ahead.

V. CONCLUSIONS

This paper presented different mechanisms to the path following problem for multipods. We formalized gaits and introduced virtual odometry to abstract from the respective leg configuration. Slippage detection and compensation is used to map planned trajectories to movement commands that are executed more precisely. We compute regulation trajectories with the help of efficient trajectory planning already used for long-range path planning to the final target. We also suggest micro regulation in case when some non-holonomic constraints can be relaxed.

The look-ahead distances currently are based on the developer's experience and experiments. Whereas small distances may lead to instabilities, larger distances increase the time to meet the planned path and moderately increase the cost value, thus are less critical. However, in the future we also want to make the ahead-distance as part of the controllable state.

REFERENCES

[1] J. Roth, “Regulated Walking for Multipod Robots”, ADAPTIVE 2019 – The Eleventh IARIA International Conference on Adaptive and Self-Adaptive Systems and Applications, May 5-9, 2019, Venice, Italy, 15-20

[2] D. Dacic, D. Nesic, and P. Kokotovic, “Path-following for nonlinear systems with unstable zero dynamics”, IEEE Trans. Autom. Control, Vol. 52, No. 3, 2007, pp. 481–487

[3] A. Morro, A. Sgorbissa, and R. Zaccaria, “Path following for unicycle robots with an arbitrary path curvature”, IEEE Trans. Robot., Vol. 27, No. 5, 2011, pp. 1016–1023

[4] P. Walters, R. Kamalapurkar, L. Andrews, and W. E. Dixon, “Online Approximate Optimal Path-Following for a Mobile Robot”, 53rd IEEE Conference on Decision and Control December 15-17, 2014. Los Angeles, California, USA

[5] S. Blažič, “A novel trajectory-tracking control law for wheeled mobile robots”, Robotics and Autonomous Systems 59, 2011, pp. 1001–1007

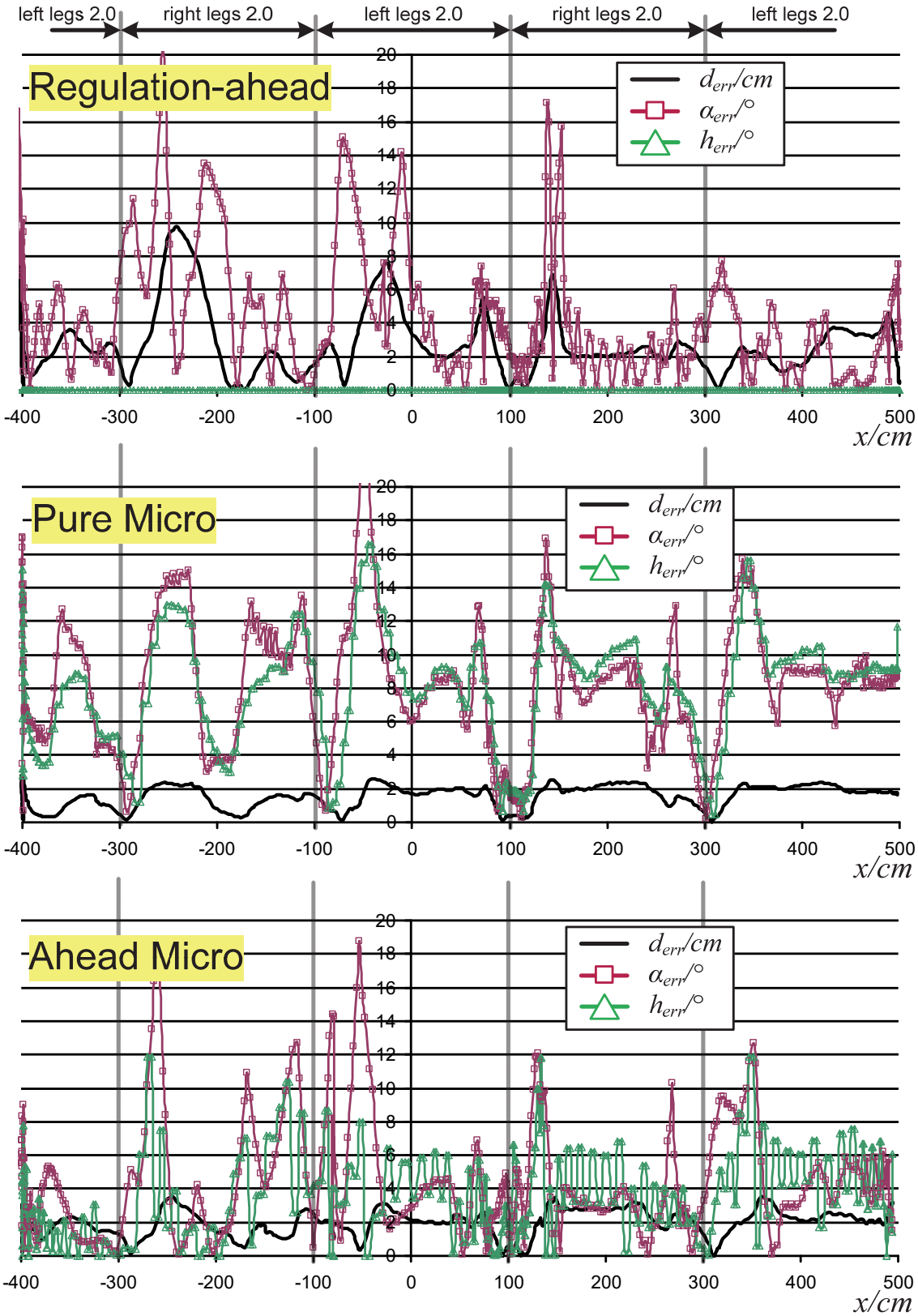


Figure 18. Detailed test run results

[6] R. W. Brockett, “Asymptotic stability and feedback stabilization", in R. W. Brockett, R. S. Millman, and H. J. Sussmann, (eds.), Differential geometric control theory, Birkhauser, Boston, 1983, pp. 181–191

[7] K. Kanjanawanishkul, M. Hofmeister, and A. Zell, “Path Following with an Optimal Forward Velocity for a Mobile Robot”, Elsevier IFAC Proceedings Volumes, Vol. 43, No. 16, 2010, pp. 19–24

[8] J. E. Normey-Rico, J. Gómez-Ortega, and E. F. Camacho, “A Smith-predictor-based generalised predictive controller for mobile robot path-tracking”, Control Engineering Practice 7(6), 1999, pp. 729–740

[9] J. J. E. Slotine, “Sliding controller design for nonlinear systems”, Int. J. Control, 40, 1984, pp. 421–434

[10] J.-M. Yang and J.-H. Kim, “Sliding Mode Control for Trajectory Tracking of Nonholonomic Wheeled Mobile Robots”, Proc. 1998 IEEE International Conference on Robotics and Automation

[11] T. I. Fossen, K. Y. Pettersen, and R. Galeazzi, “Line-of-Sight Path Following for Dubins Paths With Adaptive Sideslip Compensation of Drift Forces”, IEEE Trans. on Control Systems Technology, Vol. 23, No. 2, March 2015

[12] M. S. Wiig, W. Caharija, T. R. Krogstad, and K. Y. Pettersen, “Integral Line-of-Sight Guidance of Underwater Vehicles Without Neutral Buoyancy”, Elsevier, IFAC-Papers Online, Vol. 49, No. 23, 2016, pp. 590–597

[13] L. E. Dubins, “On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents”, American Journal of Mathematics, Vol. 79, No. 3, 1957, 497-516

[14] L. Ojeda, D. Cruz, G. Reina, and J. Borenstein, “Current-Based Slippage Detection and Odometry Correction for Mobile, Robots and Planetary Rovers”, IEEE Trans. on Robotics, Vol. 22, No. 2, April 2006

[15] C. C. Ward and K. Iagnemma, “Model-Based Wheel Slip Detection for Outdoor Mobile Robots”, IEEE Intern. Conf. on Robotics and Automation Rome, Italy, April 10-14 2007

[16] H. Deng, G. Xin, G. Zhong, and M. Mistry, “Gait and trajectory rolling planning and control of hexapod robots for disaster rescue applications”, Robotics and Autonomous Systems, 2017, pp. 13–24

[17] R. Wallace, A. Stentz, C. E. Thorpe, H. Moravec, W. Whittaker, and T. Kanade, “First results in robot road-following”, Proc. of the 9th Intern. Joint Confe. on Artificial Intelligence (IJCAI ’85), Vol. 1 , Los Angeles, Calif, USA, Aug. 1985, pp. 66–71

[18] S. Choi, J. Y. Lee, and W. Yu, “Comparison between Position and Posture Recovery in Path Following”, 6th Intern. Conf. on Ubiquitous Robots and Ambient Intelligence (URAI), 2009

[19] T. M. Howard, R. A. Knepper, and A. Kelly, “Constrained Optimization Path Following of Wheeled Robots in Natural Terrain”, in O. Khatib, V. Kumar, and D. Rus (eds.) Experimental Robotics. Springer Tracts in Advanced Robotics, Vol 39. Springer, 2008

[20] G. C. Haynes and A. A. Rizzi, “Gait Regulation and Feedback on a Robotic Climbing Hexapod”, Robotics: Science and Systems, August 16-19, 2006, University of Pennsylvania, Philadelphia, USA

[21] G. C. Haynes, “Gait Regulation Control Techniques for Robust Legged Locomotion”, PhD Thesis CMU-RI-TR-08-19, CMU, Pittsburgh, May 2008

[22] J. Roth, “A Viterbi-like Approach for Trajectory Planning with Different Maneuvers”, 15th International Conference on Intelligent Autonomous Systems (IAS-15), June 11-15, 2018, Baden-Baden, Germany, pp. 3–14

[23] J. Roth, “Robots in the Classroom – Mobile Robot Projects in Academic Teaching”, Innovations for Community Services: 19th International Conference, I4CS 2019, Wolfsburg, Germany, June 24-26, 2019, 39-55

[24] J. Roth, “Systematic and Complete Enumeration of Statically Stable Multipod Gaits”, Vol. 12, No. 4, 2018, 42-50, DOI: 10.14313/JAMRIS_4-2018/24

[25] J. Roth, “A Novel Development Paradigm for Event-based Applications”, Intern, Conf. on Innovations for Community Services (I4CS), Nuremberg, Germany, July 8-10, 2015, IEEE xplore, 69-75

[26] J. W. Gibbs, “Elements of Vector Analysis”, New Haven, 1884

[27] A. Cayley, “The collected mathematical papers of Arthur Cayley”, I (1841-1853), Cambridge University Press, 332-336, 1889

ERP Systems in Public Sector Organization: Critical Success Factors in African Developing Countries

Marie-Douce Primeau
Management and Technology Department
École des Sciences de la Gestion (ESG), UQAM
Montréal, Canada
primeau.marie-douce@uqam.ca

Marie-Pierre Leroux
Management and Technology Department
École des Sciences de la Gestion (ESG), UQAM
Montréal, Canada
leroux.marie-pierre@uqam.ca

***Abstract*— In the wake of budget restriction and increased pressure for transparency and accountability, more and more public sector organizations have opted to implement enterprise resource planning systems. Public sector organizations of developing countries have also followed this trend, pressured not only by the demands of accountability and efficiency from their own citizens but also from the multilateral and bilateral development agencies that fund a majority of the development projects and programs that they deliver. Enterprise resource planning is also seen as a way to foster organizational transformation, though best practices adoption and process harmonization. Yet, success rate of enterprise resource planning systems implementation, adoption, as well as their perceived results are less then optimal. This paper aims to explore the critical success factors in the implementation of an enterprise resource planning system in the context of public service organization in African developing countries. The results aim to guide practitioners and decision-makers with tools to increase the chances of success of these initiatives.**

Keywords- Enterprise Resource Planning – ERP; public sector organizations; Critical Success Factors – CSF; developing countries.

I. INTRODUCTION

An increasing number of public sector organizations (PSO) has opted to implement enterprise resource planning (ERP) systems [1]. This trend is also followed by developing countries, pressured not only by the same demands from their own citizens but also from the multinational and binational bilateral funding development agencies.

ERP system implementation is still in its early stages in developing countries, with Asia-Pacific and Latin America accounting for most of its expansion, and Africa trailing behind [2]. Yet, today it is estimated that developing countries account for 10% of all ERP sales [3].

In North America and Europe, the private sector is the main client of ERP systems. In developing countries, ERP are mainly deployed in large organizations, rather than in SMEs. The public sector being the largest employer in developing countries [4], the main proportion of ERP systems is implemented in PSO. This specificity adds an additional level of complexity to an already complex project, since funding usually comes in part from external single or multiple donors, with their own interests in the project, and their own procurement, management and monitoring processes. Success rate of ERP systems implementation, adoption, as well as their perceived results in PSO in developing countries are less then optimal. Yet, little research has been undertaken to understand the specific Critical Success Factors (CSF) of the implementation process of ERP in PSO in developing countries.

Based on secondary data analysis of CSF collected through four professional workshops with key stakeholders, this paper aims to explore this gap. The paper is structure as followed: Section II presents the main dimensions of an ERP systems and draw some insights specific to PSO in African context. Section III presents the methodology of this paper, while Section IV presents the main results. Section V reviews the discussion, before presenting the conclusion in Section VI.

II. CONTEXT

In this section, we will define the main terms used in this paper such as ERP, PSO and developing countries; describe the reasons why PSO would implement ERP systems; and explore main CSF in ERP systems implementation, both in general and specific to PSO in developing countries.

A. What is an ERP?

An ERP system is an "adaptable and evolutive software system that supports real-time and integrated management of a majority – if not all – processes of an organization" [5, p. 70]. ERP systems are an integrated, modular, customizable and uniform (database, management and interface) software [6][7].

ERP systems are highly complex [8]. Marnewick and Labuschagne [8] postulate that ERP systems can be conceptualize as a combination of four main components: Software (Product), Process Flow (Performance), Change Management (Process) and Customer Mindset (People; Figure 1 below). All four components are implemented through a Methodology, which underlines each ERP life-cycle phases (pre-implementation, implementation and post-implementation phases [9]).

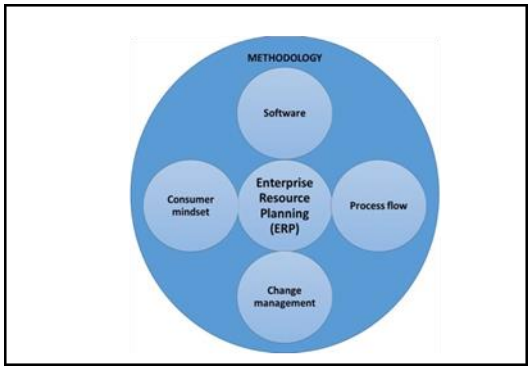


Figure I. Conceptual model for Enterprise Resource Planning (ERP), Marnewick and Labuschagne [8].

Conceptual model components: The Software component refers to the ERP product itself, such as its main features, choice of interface, and other technical aspects, as well as its development, testing and troubleshooting. The Process flow component refers to the way the different ERP modules flow within and between them. This includes both the processes themselves and the data they store and process.

The Customer mindset component refers to the need for internal stakeholder management at the user, team and organizational levels. Then, the change management component covers all factors pertaining to the planning, managing and controlling of changes. Change management is divided in four subcomponents, namely: user attitudes changes, project changes, business process changes, and system changes. Lastly, Methodology refers to the "systematic approach to implement an ERP system" [8, p.153]. All together, these components help better approach ERP system's complexity.

B. *Why would PSO want to implement an ERP system?*

PSO consists of "governments and all publicly controlled or publicly funded agencies, enterprises, and other entities that deliver public programs, goods, or services", and exists at any level – international, national/federal, regional or local) [4].

Public and private sectors have "different goals and motives and are governed by somewhat different principles, with unique groups overseeing their actions and procedures" [10]. Organizations in the private sector have "more freedom to operate, while public organizations are governed by laws, rules, traditions, and structural bureaucratic checks and balances" [10].

Although very different, benefits sought during ERP system implementation seem consistent among public- and private-sector organizations [11]. These benefits include improvements in:

- **Financial performance:** improves financial management, creates value, maximizes investments, and reduces costs;
- **Functional performance:** increases productivity, quality of services, and functional efficiency, improves management of resources, enables automation of operational procedures, eliminates redundant data and operations, and reduces cycle times;
- **Organizational performance**: increases organizational performance, enables the centralization and delocalization of maintenance services, increases adaptability, facilitates harmonization around best practices, enhances support to organizational activities, and changes nature of work in various units and departments;
- **Communication management:** centralizes and harmonizes information, improves management and organization of internal and external information flux, and improves security and information access management;
- **Internal audit, monitoring and control:** improves controls and institutional accountability, enhances organizations regulatory compliance, achieves accuracy in management information system, enables real-time access to performance information, which in turn fosters better strategic analysis and decision [6] [12] [13].

Furthermore, a study on the impact of ERP systems in small and midsized PSO suggests that implementing an ERP system helped PSO improve services to customers and suppliers while enhancing knowledge of primary users and increasing shareholders confidence in organization [13]. With all those potential benefits, we have to ask: why are not more PSO implementing ERP systems?

C. *Is ERP implementation in PSO sucessful?*

As discussed above, ERP system implementation can enhance benefits for PSO. Nevertheless, ERP system implementation can be cost and time consuming [14]. As example, the cost of ERP implementation in the United Nations (UN) organizations is estimated at 712 million United States Dollar (USD). This does not include recurring maintenance costs (at least 66 million USD per year), nor the off-budget associated costs (between 86 and 110 million USD per year).

Furthermore, failure rate, both in private and public organization, is high. The 2016 ERP Report [15] states that less that 10% of all ERP projects sampled in 2015 were implemented on time, within budget and in respect to the planned scope. More than a third (35%) was stopped or (indefinitely) differed. The remaining 55% were completed with an average of 178% cost and 230% schedule overruns. In fact, ERP implementation projects lasted 1 to 3 years, with an average of 21 months, while most projects had been planned around an 8-14 months' timetable.

Although data on the subject is scarce, ERP systems implementation failure rate in PSO in developing countries is believed to be even higher. In his study of ERP implementation in Egyptian organizations, Abdelghaffar [16] argued that 75% of ERP implementation attempts can be classified as failures. Another study found schedule overruns in 67% and cost overruns in 33% of all ERP implementation projects in United Nations organizations

[12]. Reasons frequently mentioned to explain these schedule overruns were: changes in project scope; delays in personalization of software; users' resistance to change, delays in data conversion, changes in initial project strategy, and redefinition of operating procedures. As for cost overruns, they were attributable mainly to unplanned personalization costs; inadequate definition of functional needs; unforeseen delays in the implementation process, and unrealistic cost estimation planning. No data was found on ERP implementation success in African developing countries, even if failure rates are thought to be higher than in developed countries [6].

D. Are all PSO the same? or How do PSO from developing countries differ from PSO from African developed countries?

United Nations divides countries into two categories: developed and developing countries. This classification is mainly based on economic indicators and indices such as Gross Domestic Product (GDP), Gross National Product (GNP), per capita income, unemployment rates, industrialization and standard of living [17]. The developing countries categories include both developing and least developed countries, most of which are in Africa.

Contrary to developed countries, most PSO in African developing countries are funded (partly of entirely) by external funding. Funds come mainly from multi donors/multilateral aid agencies in the context of national strategy to capacity building. In exchange for grants or concessional loans, beneficiary countries are expected to report their results, and be accountable. In this respect, all funded initiatives, whether in the form of technical assistance or capacity building projects, in all sectors, including governance, are required to be designed, executed and evaluated under a results-based management approach. It is indeed under the auspices of these major capacity building programs for public administrations that ERP projects have often been imposed as a way to increase transparency and guarantee accountability [6]. In this vein, local participation in the project has been a key message to increase ownership of public bodies in developing countries. For years, the participation of beneficiaries in the process and the management of the funds allocated to them has been part of participatory approaches, which stipulates that local participation in donor-funded initiatives becomes an essential ingredient in ownership.

However, considering the important costs – both financial, social and political – associated to ERP implementation failures in PSO in African developing countries, it is important to understand the CSF that could hinder or facilitate this process.

E. What are the CSF in ERP systems implementation in PSO in African developing countries?

In order to support organizations in their implementation efforts, practitioners and researchers have come up with CSF that facilitate or hinder implementation. CSF are defined as "factors needed to ensure a successful ERP project" [18]. This includes both factors that facilitate and hinder the implementation of an ERP system. These factors vary according to the nature and environment of the organization [19]. Yet most research on ERP success factors have been done in developing countries, in the context of private-sector organizations.

Through their literature review of CSF in ten different countries/regions, Ngai, Law and Wat [19] identified eighteen CSF, with more than 80 subfactors for the successful implementation of an ERP. The CSF are: appropriate business and IT legacy system; business plan/vision/goals/justification; business process reengineering; change management, communication; data accuracy; ERP strategy and implementation; ERP project team; ERP vendor; monitoring and evaluation performance; organizational characteristics; project champion; project management; software development, testing, and troubleshooting; top management support; fit between ERP and business/process; national culture; and country-related functional requirements [19]. This typology has been used by other scholars to guide their analysis of the influence of CSF in phases of an ERP implementation process [20] [21].

In the last years, few studies have tried to identify CSF specific to ERP implementation in PSO of developing countries.

In its assessment of ERP implementation projects in its organizations, the United Nations identified eleven CSF, namely: project planning and software selection; governance of the project, risk management, change management, project team, end users training and assistance; ERP system hosting and infrastructure; data conversion and systems integration, ERP upgrade, and project audit [12].

Another study from the World Bank identified eight CSF from its experience implementing ERP systems, namely: capacity building and training, close supervision and control from the donor agency, favorable political context and leadership; pre-existing favorable environment (IT, HR, Accounting); adequate preparation and clear conception; good project management and coordination, and external environment factors [22]. It also identified main failure factors, which were: inappropriate training/education of project teams; institutional/organizational resistance; inadequate project preparation and planning; complex conception/high number of procurements; organizational structure adapted to integration efforts; inadequate IT infrastructure; absence of leadership/engagement and ambiguous attitude of authorities regarding implementation; inappropriate technology; inadequate project coordination; and external factors (political troubles, natural disasters). These failure factors are consistent with other studies on ERP implement issues in developing countries [14] [23] [24].

These studies offer some insight on perceived CSF in ERP implementation from the point of view of donor agencies. Yet, these highlight the need to further explore the Critical Success Factors (CSF) in the implementation of an ERP system in PSO in African developing countries, in hope to give practitioners and decision-makers tools to increase

the chances of success of these initiatives. This paper will try to address this gap.

III. METHODOLOGY

This work uses secondary data collected through professional workshops with key stakeholders that have direct experience either in the planning, managing or implementing of an ERP in PSO in developing countries. A description of the initial data collection process and methods, as well as an overview of the data analysis techniques and conceptual model used for secondary data analysis follows.

A. Data collection – primary data

Primary data was collected through four 1 ½- 2 hours professional workshops on successful ERP implementation. In total, 140 participants took part in the workshops. The workshops took place in Abidjan (Ivory Coast), Rabat (Morocco) and Marrakech (Morocco). Participants from workshops 1 and 3 were all locals, while participants from workshop 2 were mostly locals, and all participants except two from workshop 4 were from outside of the country, namely from other West African countries. In total, 104 participants gave out their information contacts to organizers, for a 74% answer rate. Out of these, 62.5% of participants came from Ivory Coast, 20.5% from Morocco, 4% from Guinea, 3% from Burkina Faso, 3% from Benin, 2% from Mauritania, 2% from Senegal and 1% from Mali.

The following subsection offers an overview of the composition of each of the workshop groups.

- **Workshop no. 1:** 15 participants from a multilateral development bank institution working as Task team Leaders, Procurement and Monitoring and Evaluation. Specialists, and Managers. Languages: English and French.
- **Workshop no. 2:** 85 participants from public and parapublic organizations. Participants worked as directors, project or program managers, procurement or monitoring and evaluation sectors on bilateral or multilateral initiatives. Two came from the academia. Language: French.
- **Workshop no. 3:** 26 participants from public organization sector or project and programs funded through bilateral or multilateral development aid. Languages: French and Arabic.
- **Workshop no. 4:** 14 participants from West Africa working as either project or program managers or Monitoring and Evaluation Specialists on bilateral donors or multilateral projects or programs. Language: French**.**

The diversity within the different groups was one of the main difficulty / challenges encountered by the workshop facilitators (English/French/Arabic languages, professional status, type of organizations, and number of participants per session). To increase participation, reduce cultural barriers, provide a safe climate to exchange and create cohesion between participants of the workshops, facilitators used World Café as a data collection method.

World café is a collaborative approach that aims "to engage [participants] in constructive dialogue around critical questions, to build personal relationships, and to foster collaborative learning [25, p.28]", and helping creative new ways to address problems emerge from the initiative. Simple and flexible, the approach can be used both in small and large heterogeneous groups to foster open dialogue and collaboration [26].

World café follows seven integrated design principles, namely:

- □ Set the context;
- □ Create a hospitable space;
- □ Explore questions that matter;
- □ Encourage everyone's contribution;
- □ Connect diverse perspectives;
- □ Listen together for patterns and insights;
- □ Share collective discoveries [22].

At the end of each of the workshops, participants drafted a list of factors that facilitated and hindered the implementation of an ERP. All entries of the four lists were then combined by the facilitators. This final compilation was sent to participants in the conference proceedings by the workshops organizers. These conference proceedings are the basis of our analysis.

B. Data analysis

All entries of the conference proceedings were analyzed and combined through thematic analysis [27]. To facilitate understanding, subthemes were then organized using a modified version of Marnewick and Labuschagne [8]'s ERP Conceptual Model. This modified version includes all four main components (Software, Process Flow, Change Management, Costumer Mindset), Methodology, and adds a last component - external environment. This component was added to consider the influence of national culture [19] and other macroeconomic factors pertaining to the implementation of ERP systems in African developing countries. The ERP project financing also falls under this category, as it has a major impact on ERP implementation in developing countries [12].

IV. RESULTS

The following section presents our results, namely the CSF identified and categorized, using the adapted conceptual model. To facilitate understanding, results are presented per components, namely: Software, Process flow, Customer mindset, Change management, Methodology, and External environment. In total, forty-one CSF were identified through this process (see Table I in the appendix).

A. Software

In total, five CSF were identified by participants for the Software component, namely: participatory software development, testing and troubleshooting; fair and balanced ERP vendors/suppliers' relationships; country-related functional requirements; adequate ERP infrastructure and hosting; and sufficient IP maturity of organizations.

Participatory software development, testing and troubleshooting: participants underlined the importance of the choices made through theses phases, and the need for user participation in the process to facilitate adoption. They highlighted the difficulties associated with the fact that these steps are often outsourced, without real inputs from directly implicated PSO stakeholders (e.g., users, M&E specialists, etc.). Furthermore, the lack knowledge transfer to local IT teams throughout the development, testing, and troubleshooting phases complicate not only maintenance, but hinders the adaptation of the software to PSOs needs.

Fair and balanced ERP vendors/suppliers' relationships: Participants highlighted that the absence of local vendors gives disproportionate power to international vendors, thus hindering optimal selection of ERP systems by PSO. Furthermore, vendors seemed reluctant to adapt their products to PSOs particular needs, knowing that they will have to buy their products anyway.

Country-related functional requirements: Participants also discussed the fact that ERP often did not meet their specific PSO requirements, e.g., integration of performance indicators at the result level, reporting formats that do not fit the donor requirements, etc. In many cases, PSO needed to combine the ERP with other monitoring tools (e.g., Excel sheets and MS Project) to fulfill their monitoring requirements.

Adequate ERP infrastructure and hosting: More and more ERP systems are cloud-based. Because of the lack of access to basic amenities in many parts of African countries, many ERP options are not feasible. ERP hosting is also a problem, not only because of security issues but also because of limited access to electricity.

Sufficient IT maturity of organizations: Participants also underlined the low IT maturity in most African PSO, which hinders their ability to facilitate ERP implementation, and to maintain the system adequately. This situation furthers their dependence on ERP vendors, and limits appropriation of the system by local IT teams.

B. Process flow

The Process flow component includes two subcategories: Process and Data. In total, seven CSF were identified by participants for the Process flow component.

a. Process

In total, three CSF were identified by participants for the Process subcomponent, namely: fit between ERP and an organization's procedures; harmonized practices, procedures and processes; and good communication management processes.

Fit between ERP and an organization's procedures: PSO in developing countries, because of their funding and organizational structure, have specific procedures (e.g., burdensome administrative and procurement procedures, strict monitoring and evaluation requirements, etc.). ERP systems are created around private-sector (occidental) best practices. Therefore, the product offered is often than not difficult to adapt to African PSO's needs.

Harmonized processes and procedures: ERP systems aim to limit the possibility of errors by limiting the number of times a same information has to be entered in the system. Yet, because of the lack of harmonized procedures, users still have the obligation to enter information on multiple software.

Good communication management processes: participants highlighted the need for clear and effective communication and information management processes, for example in sharing management's plan, in order to maximize the probability of successful implementation.

b. Data

In total, four CSF were identified by participants for the Data subcomponent, namely: efficient data quality control, good data collection and methods, solid data management practices, and clear data conversion plan and management.

Good data collection processes and methods: to populate an ERP, you need data. Participants discussed the need for an effective monitoring and evaluation (M&E) system that promotes good data collection processes and methods. On the other hand, unrealistic frequency of collection or level of detail of data requested was believed to hinder the support for the ERP implementation project.

Efficient data quality control: once you have data to populate your ERP, you have to trust it. Participants highlighted the need to have in place efficient quality control processes, to ensure data reliability. Ultimately, data of questionable quality was perceived to be associated with a reluctance from users and other stakeholders to adhere to the ERP implementation project.

Solid data management: good and solid data management was considered at the core of ERP implementation process. This included not only processes to insure data management as a hole, but also addresses the topics of data security, access, and traceability.

Data conversion plan and management: data conversion, meaning planning, managing and controlling of prior data integration in the ERP, was seen as a core component of an ERP implementation project by participants. Yet, many highlighted the lack of actual planning around this activity. Participants suggested the need for standardized guidelines and processes around data conversion that would cover: which data to conserve and, which to drop; how far back should PSOs go in converting data; what format to choose; who is responsible to integrate this data, etc.

C. Change management

The Change management component can be divided into four subcomponents, namely: user attitude, project change management, business process change management, and system change management. In total, nine CSF were identified by participants in the Change management component.

a. User attitude

Participants identified three CSF pertaining to user attitude management, namely: Need for communication, Need for training and education, and User active participation in ERP implementation.

Effective communication of the change to users:

Information and effective communication with users are crucial tools to manage expectations, facilitate appropriation and ease tensions with users. This process should start at conception, and continue throughout the project, and be planned carefully and strategically.

Adequate training of users: Often, training is seen as the last milestone before transfer to operations. Yet, participants highlighted the need for training to start earlier in the process, since it often uncovers problems or incongruities with the ERP system and its application to their everyday work. When done to late in the process, project often do not have the resources (and time) to redo the work, therefore delivering a product that does not fit users' needs.

Active participation of users: Users needs differ from other stakeholders, and can be misunderstood. As for communication, users' participation is essential in an ERP's appropriation and future use, and should be planned throughout the implementation project's life-cycle.

b. Project change management

Participants identified one CSF pertaining to Project change management, namely: effective change control management processes and procedures.

Effective change control management processes and procedures: Changes are inevitable in projects. Yet, the absence of a formalized and effective change control management, and clear procedures to support it, made it difficult for ERP implementation projects to stay on track. At opposite, effective change control management seems to have helped project managers to limit unnecessary changes to the project scope, by giving them the tools to answer to stakeholders' pressures and change demands that might fall outside the intent of the project.

c. Business processes change management

Participants identified two CSF pertaining to Business process change management, namely: harmonization of practices and processes, and assessment of best practices.

Harmonization of practices and processes: The importance of understanding all business systems, policies and institutional procedures to ensure better alignment with ERP system functions. For example, procurement requirements and local vs. international accounting standards are not compatible with the system's data collection procedures.

Assessment of best practices: Participants expressed the need to know more about best practices before making any changes. They say they want to be informed of best practices in the African context in order to continue efforts towards continuous improvement and institutional capacity building.

d. System change management

Participants identified three CSF pertaining to System change management, namely: management of interests; communicate change throughout the organization; and plan and manage corporate culture change.

Management of interests: participants highlighted the importance of targeting the expectations of different users and other stakeholders, not all of whom have converging interests.

Communicate change throughout the organization: All required changes should be communicated in advance. The need to put in place communication procedures to promote acceptance and ownership of changes throughout the implementation of the system.

Plan and manage corporate culture change: Participants also mentioned the importance of matching the organizational culture with the desired properties and functions of the ERP system. For example, a shared values charter at the beginning of the project was mentioned as an element in an organizational change management plan that accompanies process re-engineering.

D. Customer mindset

The Customer mindset component includes three subcategories, namely: User mindset, Team mindset, and Organizational mindset. In total, fourteen CSF were identified by participants for the Customer mindset component.

a. User mindset

In total, three CSF were identified for the User mindset subcomponent, namely: users' attitudes/openness to change, adequate technical level of competencies and knowledge of users, and access to training.

User attitude/ openness to change: Openness or, on the contrary, resistance to change was systematically highlighted as a major factor influencing success of ERP implementation.

Adequate technical level of competencies and knowledge of users: users need to have sufficient knowledge of computers systems and IT competencies to be able to not only feed data but also use efficiently the ERP system.

Access to training: ERP systems modifications and upgrades are inevitable; so are new hires or transfers in teams using ERP systems modules. Therefore, users need access not only to initial but also to continuous training to be able to stay current with the latest development of the ERP system.

b. Team mindset

In total, five CSF were identified for the Team mindset subcomponent, namely: adequate team competencies, team composition, stability of teams / low attrition rate, good collaboration, and leadership.

Adequate team competencies: Participants highlighted the need for a multidisciplinary and diversified team that addresses both the IT component, but also the organization change management facets of an ERP implementation project.

Team composition: participants also discussed the influence of differences of status/treatment and employment on the team mindset.

Stability of teams/Low attrition rate: In some African countries, PSO's employment conditions (such as salary, insurance, etc.) makes private employment more attractive in sectors such as IT.

Good collaboration: collaborative relationships between team members are essential to navigate the complexity and problem diversity of ERP implementation. This includes:

good team work, respect between co-workers, and collaboration.

Leadership: ERP projects being complex, participants also highlighted the need for leadership inside the team, for example to avoid being lost in multiple stakeholders demands. Strong management skills from team leaders was also put forward as a CSF in ERP implementation.

c. Organization mindset

In total, seven CSF were identified for the Organization mindset subcomponent, namely: prior experience in ERP/major IT project implementation, change management competency, organizational support/commitment, presence of a champion, shared vision, mission and organizational goals, stakeholders' ownership of the project; and need for real-time information.

Prior experience in ERP or major IT project implementation: participants highlighted PSO experience in implementing similar projects (in form or complexity) as a CSF of ERP implementation success.

Change management competency: Participants systematically identified PSOs' change management abilities or competencies as CSF for ERP implementation success.

Organizational commitment: ERP systems implementation include not only a prior preparatory phase, the project phase itself, but also maintenance and upgrades. Furthermore, with a 5-8 year product life-cycle, ERP implementation can be seen as a long term commitment for PSOs, that will require both funding, adequate staffing and logistics.

Presence of a champion: participants highlighted the importance of having a champion. This person needs to be part or linked to high management of PSO, and have sufficient power within the organization.

Shared Vision, mission and organizational goals: ERP are useful tools to operationalize an organization's strategy. Yet, to be able to perform, participants highlighted the need for shared vision, mission and organizational goals. This include: mission and vision definition, communication and appropriation by stakeholders.

Stakeholders' ownership of project: all stakeholders need to feel implicated in the project, and have a sense of responsibility toward the success of the ERP implementation project – and its utilization.

Need-driven endeavor: To be successful, participants highlight that the ERP must be understood as a mean to an end, such as the need for real-time information.

E. Methodology

In total, two CSF were identified by participants for the Methodology component, namely: good project management, and clear ERP implementation strategy.

Good project management: Participants stressed the importance of good project management in ERP implementation, namely the need for clear planning, project division in multiple steps; realistic performance demands and deadlines, collecting of lessons learned; planning of implementation costs and maintenance.

Clear ERP implementation strategy, and its communication to stakeholders, were also seen prerequisite for ERP implementation success.

F. External environment

In total, three CSF were identified by participants in the External environment component, namely: fit with national culture and values; balanced donor-recipient relations; and adequate local infrastructure.

Fit with national culture and values: the participants mentioned the lack of coherence between some habits and customs and the purposes of a well-established ERP. ERPs aim to foster transparency and accountability in public projects, therefore supporting the fight against corruption in PSO.

Balanced donor-recipient relations: More often than not, donors were not only (openly or not) the instigators of the ERP implementation project, but also guided the choice of vendors/suppliers. "Give and take" in the needs of both donors and recipients was seen as a CSF of ERP implementation success.

Adequate local infrastructure: Access to electricity, telecommunications and Internet remains problematic in many African countries, especially when outside urban agglomerations [28], though significant improvements have been made in recent years. This has a major impact not only on ERP implementation but adoption by users.

V. DISCUSSION

As mentioned, ERP systems implementation projects aims to the achievement of organizational benefits [9], regardless of the nature of organizational, private or public activities. Yet, our results suggest that the benefits sought in terms of financial, organizational, communication and evaluative performance in African PSOs are limited by various barriers identified in this study. Furthermore, our results highlight the specific nature of ERP systems implementation in PSO in African countries, and the need to better understand how these specific CSF influence ERP implementation success in these context.

Software dimension: Yet, of the five CSFs in the Software dimension, three (participatory software development/testing/troubleshooting, fair and balanced ERP vendors/suppliers' relationships, and country-related functional requirements) seem to be specific to PSOs in African developing countries.

Participation stands out as the first CSF in the software dimension. Participatory approach remains one of the methodologies advocated as a key success factor in international development projects. This is also what was identified in Poonam and Agarwal's ERP study [29]. Yet this participatory approach seems difficult to achieve considering the unequal power balance between stakeholders in international development.

The power balance between the vendors and African PSOs is an unequal relationship, where PSOs do not have

all the knowledge to make informed choices, nor the power to influence the outcome of the decision process [30] [31]. This limits the ability of PSOs to introduce country-specific requirements in the contract negotiations.

Furthermore, the funding agency have been known to have direct interest with vendors and ERP service providers, biasing the selection processes. Even when this is not the case, the choices must meet the requirements set out in the loan agreement, which limits the options to use local suppliers. This situation complicates the relationship between the vendors or the supplier, and the customer, in this case the PSO, since the contractual basis, or tacit contract, includes three parties. Suppliers are often selected for their track record in large Western organizations [31], which do not have the same organizational maturity as understood by Western standards [32]. New African-based competitors – although their products are far from optimal at this moment – hopefully will change this dynamic, and might foster a more balanced relationship between vendors and PSOs. Surprisingly, these two CSFs were not identified in the study conducted by the World Bank Group [22].

Process dimension – Data sub-dimension Data quality and reliability is essential to any ERP implementation project. Yet, what makes this CSF specific to African developing PSOs is the scale of the M&E necessary to populate the ERP. It is important to keep in mind that in African developing context, one of the main drivers of ERP implementation project is to support PSOs in their efforts to provide proof of results in light of funding received. This is achieved through the implementation of a results-based management system that encompasses the monitoring of outcomes and results indicators [33]. These indicators are collected either through project and program-funded data collection, or through national statistical data collection agencies – both of variable and questionable quality in developing countries [34]. This highly contrasts the context of private organization, where most data integrated into the ERP is internally collected, and quality can be more easily controlled.

In these contexts, ERP implementation and deployment initiatives are often accompanied by a myriad of institutional capacity building measures, such as donors finance reform initiatives [29]. In exchange for support in public financial management, donors will often include conditionalities such as demands for improvements in technological infrastructure in their funding agreement. These will in turn influence the ERP implementation project, therefore adding a level of complexity for African PSOs that other organizations do not have.

Furthermore, although several CSFs had been identified in previous studies [10] [17] [18], a closer reading of the results provides some nuances. For instance, «Team composition», «collaboration», «leadership» and «competencies» were found to be CSF in all studies [10] [17] [18]. However, the way they materialize differs. To illustrate this, let us take the CSF «Team composition» as an example.

All studies agree on the importance of building teams that are diversified in terms of skills, experience and abilities, with good intra and inter collaborative skills [10] [17] [18]. The importance of collaboration is not specific to the ERP implementation project. A recent study identified this factor as central to the equation leading to capacity building in so-called developing countries [35]. However, our results attempt to demonstrate that the retention of employees assigned to ERP implementation projects is problematic. Several factors explain this situation, including the high rate of absenteeism and the lack of incentives for public servants [36]. The stability of teams is often compromised, and consultants, who are highly solicited in these types of complex projects, accentuate the motivational problems of government employees. In most cases, consultants assigned to ERP implementation projects, often referred to as "technical assistants", receive much higher compensation than civil servants [37]. This gap is even more acute when the technical assistant comes from an OECD member country, for example. This situation, perceived as unfair by local team, has negative effects on the dynamics of project teams and their performance. ERP project teams in developing countries are a combination of consultants, who are often lent by the PSO themselves (not always for their competencies), and that are paid in a day what the rest of the teammates will sometimes do in a month. The apparent unfairness in the treatment of team members, although important, may be accompanied by other elements that should be addressed.

Another example of CSF's specificity is the «Organizational commitment». In all studies, organizational commitment, such as support from top management, is seen as a major CSF for ERP implementation. The World Bank Group's study [22] also linked this CSF to the CSF labeled «suitable political environment». While the majority of the studies cited this variable as one of the most commonly identified CSFs, the underlying explanation differs from other contexts. Yet, in African PSOs, top management is often the one who benefits from the lack of transparency and accountability [38], and therefore are the main opponents of these type of initiatives [39].

Change management dimension Our results also found that CSFs relating to Change management dimension encompasses all organizational levels. However, as noted above, if the initiative comes from an external source, the different dimensions of change (user attitude, business process change and system changes) may suffer from internal support, thus limiting internal initiatives for preparing and adapting to change. Of the 9 CSFs listed under this dimension, the "assessment of best practices" and "management of interests" seem to be only found in this context. On the one hand, best practices are taken from Western organizations, which do not seem to be adapted to the contexts of African PSOs. This universalist approach shows some limitations, as previously highlighted by Hasheela-Mufeti [40]. On the other hand, «management of interest» is one of the CSFs that complicates the

management of ERP implementation projects, since many external stakeholders are involved in these initiatives.

External environment dimension Lastly, External environment dimension appears to be highly relevant in African PSO context. Our results reflect the many CSF that need to be considered in order to increase the chances of success of ERP implementation projects. The most salient refers to the level of development of public infrastructure, such as electricity and technology. Again, although the study focuses on PSOs in so-called developing African countries, levels of "development" vary between countries. Following the example of the Gapminder Institute's work [42], a large proportion of African countries are at level 2 (out of 4), which results in inadequate public services in several respects. This situation is reflected in frequent power cuts, a faulty or even non-existent Internet network in many cases. The reliability of these two types of infrastructure, in terms of access and availability, remains a major challenge [43]. The CSFs identified in this dimension are the first factors that all organizations working in international development must evaluate before even starting any ERP implementation project, regardless of its size.

VI. CONCLUSION

ERP implementation projects are often wrongly considered IT projects, when in fact they are major organizational transformation initiatives [22] that will significantly change the processes, structure, even the culture of an organization [10]. In line with current research [12], the need for training and education (both for users and project team members), top management support and multilevel change management were most cited CSFs by participants.

Team members' mindset also seems to have a major influence on ERP implementation success – and on its failure. Yet, current research has done little to study the issues specific to the dynamics of the teams in charge of implementing ERP implementation projects in the context of African countries receiving international aid, with all financial and legal complexities that it implies.

Ultimately, our results highlight that CSF' influence vary depending of many factors, such as organizational and national culture, type of implementation process chosen (one time or gradual implementation), etc. This converge with Zouagui and Laghouag's findings [44]. Yet, these specificities are rarely taken into account in ERP implementation in PSO in African developing countries projects. While this study highlights factors that seem specific to ERP implementation in West African countries (e.g., fit with values, etc.), other could be generalize to all countries that rely on international development (e.g., fair and balanced ERP vendor/buyer relationships, balanced donor-recipient relations, etc.). Still, further research is needed to better understand and conceptualize the CSF in ERP implementation in PSO in the African developing countries.

REFERENCES

[1] M.-D. Primeau, "Critical Success Factors in the Implementation of ERP Systems in Public Sector Organizations in African Developing Countries", the Ninth International Conference on Business Intelligence and Technology (BUSTECH 2019) IARIA, May 2019, pp. 21-26, ISSN: 2308-4391, ISBN: 978-1-61208-710-8.

[2] Grand View Research - GVR, ERP Software Market Analysis By Deployment (On-premise, Cloud), By Functions (Finance, Human Resource (HR), Supply Chain), By Verticals (Manufacturing & Services, BFSI, Health Care, Retail, Government Utilities, Aerospace & Defense, Telecom), by End-User (Large Enterprises, Medium Enterprises, Small Enterprises) And Segment Forecasts To 2022, May 2016. ID: 978-1-68038-669-1.

[3] A. Hawari and R. Heeks, "Explaining ERP Failure in A Developing Country: A Jordanian Case Study", Journal of Enterprise Information Management, vol.23, pp. 135-160, February 2010, doi: 10.1108/17410391011019741.

[4] S. Dube and D. Damescu, "Supplemental Guidance: Public Sector Definition", The Institute of Internal Auditors, December 2011, [Online] Available from: https://na.theiia.org/standards-guidance/Public%20Documents/Public%20Sector%20Definition.pdf. [retrieved: April 12th, 2019].

[5] S. Uwizeyemungu, "Evaluation of integrated software on organizational performance: process methodology development / L'évaluation de la contribution des progiciels de gestion intégrés à la performance organisationnelle: Développement d'une méthodologie processuelle". Doctorate dissertation (Doctorate in Business Administration) presented to the Université du Québec à Trois-Rivières, 353 p., 2008.

[6] P. M. Kengue, "ERP and integrated information system contribution in ONU driven development projects/ Contribution des ERP et des systems d'information intégrés dans les projets de développement pilotés par l'ONU: Analyse empirique de la situation au travers de trois missions d'expertise (Brazzaville, New York, Bonn) abordées sous l'angle du bricolage ciborrien", Doctorate dissertation (Doctorate in IT Systems Management) presented to the Université de Nantes, 297 p., 2015.

[7] K. Debbabi, "Cognitive and emotional determinants of the acceptability of information and communication new technologies: the case of ERP / Les déterminants cognitifs et affectifs de l'acceptabilité des nouvelles technologies de l'information et de la communication : Le cas des Progiciels de Gestion Intégrés", Doctorate dissertation (Doctorate in Work Psychology and Ergonomy) presented to the Université de Grenoble Alpes, 248 p., 2014.

[8] C. Marnewick and L. Labuschagne, "Conceptual model for enterprise resource planning (ERP)", Information Management & Computer Security, vol. 13, pp. 144-155, April 2005, doi: 10.1108/09685220510589325.

[9] S. I. Chang, G. Gable, E. Smythe, and G.Timbrell, "A Delphi examination of public sector ERP implementation issues". The 21st International Conference on Information Systems (ICIS 2000), December. 2000, pp.494-500, Available from: http://eprints.qut.edu.au/archive/00004746 [retrieved: April 10th, 2019].

[10] San Francisco University, "5 *Key Differences Between Organizations from the Public and Private Sectors*", [Online] Avaliable from; https://onlinempadegree.usfca.edu/news- resources/news/5-key-differences-between-organizations-in-the-public-and-private-sectors/, [retrieved: April 12th, 2019].

[11] J. L. Harrison, "Motivations for Enterprise Resource Planning (ERP) System Implementation in Public versus

Private Sector Organization", 2004, Electronic Theses and Dissertations 31, [Online] Available from: https://stars.library.ucf.edu/etd/31/ [retrieved: April 10th, 2019].

[12] United Nations Group, "Study of management software integrated in United Nations organizations", 2012, [Online] Available from http://undocs.org/fr/A/68/344. [retrieved: April 12th, 2019].

[13] A. R. Singla, "Impact of ERP Systems on Small and Mid-Sized Public Sector Enterprises", Journal of Theoretical and Applied Information Technology, vol. 4, February. 2008, pp. 119-131.

[14] A.T. Manga, R. A. Etoundi, and J. Zoa, "A Literature Review of ERP Implementation in African Countries", Electronic Journal of Information Systems in Developing Countries, vol.7, pp.1-20, September 2016, doi: 10.1002/j.1681-4835.2016.tb00555.x.

[15] Panorama Consulting Solution Group, "2016 Report on ERP Systems and Enterprise Software", [Online] Available from: https://www.panorama-consulting.com/wp- content/uploads/2016/07/2016-ERP-Report-2.pdf [retrieved: April 12th, 2019].

[16] H. Abdelghaffar, "Success Factors for ERP Implementation in Large Organizations: The Case of Egypt", The Electronic Journal of Information Systems in Developing Countries, vol.52, pp.1-13, December 2012 doi: 10.1002/j.1681- 4835.2012.tb00369.x

[17] United Nations Group, "Working Together: integration, institutions, and the Sustainable Development Goals: World Public Sector Report 2018", UN Division for Public Administration and Development Management, Department of Economics and Social Affairs (DPADM), New York, April 2018 [Online] Available from: http://workspace.unpan.org/sites/Internet/Documents/UNPAN 98152.pdf, [retrieved: April 12th, 2019].

[18] D. Allen, T. Kern, and M. Havenhand, "ERP Critical Success Factors: an Exploration of the Contextual Factors in Public Sectors Institutions", The 35th Annual Hawaii International Conference on System Sciences (HICSS-35), Jan. 2002, pp.3062 – 3071, ISBN: 0-7695-1435-9, doi: 10.1109/HICSS.2002.993840.

[19] E. W. T. Ngai, C. C. H.Law, and F. K. T. Wat, "Examining the critical success factors in the adoption of enterprise resource planning", Computers in Industry Journal, vol.59, pp.548-564, August 2008, doi: 10.1016/j.compind.2007.12.001.

[20] M.R. Moohebat, A. Asemil, and M.D. Jazi, "2A Comparative Study of Critical Success Factors (CSFs) in Implementation of ERP in Developed and Developing Countries", International Journal of Advancements in Computing Technology, vol. 2, December 2010, doi: 10.4156/ijact.vol2.issue5.11.

[21] J. Ram, D. Corkindale, and M.-L. Wu, "Implementation critical success factors (CSFs) for ERP: Do they contribute to implementation success and post-implementation performance?", International Journal of Production Economics, vol. 144, pp. 157-174, January. 2013, ISSN: 0925-5273, [Online] Available from: http://doi.org/10.1016/j.ijpe.2013.01.032.

[22] C. Dener, J.A. Watkins, and W.L. Dorotinsky, "Financial management information systems : 25 years experience in World Bank on what's working, and what's not", Washington DC : World Bank, doi: 10.1596/878-0-8213-8750-4, 2011.

[23] Z, Huang and P. Palvia, "ERP Implementation Issues in Advanced and Developing Countries", Business Process Management Journal, vol. 7(3), pp. 276-284, 2001,doi: 10.1108/14637150110392773.

[24] A. Asemi and M. D. Jazi, "A Comparative Study of Critical Success Factors (Csfs) in Implementation of ERP in Developed and Developing Countries", International Journal of Advancement in Computing Technology, vol. 2, pp. 99-110, December 2010, doi: 10.4156/ijact.vol2.issue5.11.

[25] C. Fouché and G. Light. "An invitation to Dialogue: 'The World Café' in Social Research Work", Qualitative Social Work, vol. 10, pp. 28-48, March 2011, doi 10.1177/1473325010376016.

[26] J. Brown and D. Isaacs, World Café Community, M. J. Wheatley and P. Senge, "The World Café Book: Shaping our Futures Through Conversations that Matter", CA: Berrett- Koehler Publishers, 2005.

[27] R. Boyatzis, "Transforming qualitative information: Thematic analysis and code development". Thousand Oaks, CA: Sage, 1998.

[28] P. Kraemmerand, C. Moller, and H. Boer, "ERP Implementation: an Integrated Process of Radical Change and Continuous Learning", Production Planning and Control, vol. 14, pp. 338-348, 2003.

[29] G. Poonam and D. Agarwal, "Critical success factors for ERP implementation in a Fortis hospital: An empirical investigation", Journal of Enterprise Information Management, vol. 27, pp. 402-423. July, 2014, https://doi.org/10.1108/JEIM-06-2012-0027.

[30] M. Dornan, "How new is the 'new' conditionality? Recipient perspectives on aid, country ownership and policy reform", Development Policy Review, vol. 35, pp. 46-63. July 2017, https://doi.org/10.1111/dpr.12245

[31] S.E. Fischer and M. Strandberg-Larsen, "Power and Agenda-Setting in Tanzanian Health Policy: An Analysis of Stakeholder Perspectives", International journal of health policy and management, vol. *5*, pp. 355–363. June 2016, doi:10.15171/ijhpm.2016.09

[32] S.O. Babatunde, S. Perera, and L, Zhou, "Methodology for developing capability maturity levels for PPP stakeholder organizations using critical success factors", Construction Innovation, vol. 16, pp. 81-110. January 2016, https://doi.org/10.1108/CI-06-2015-0035

[33] D.W. Brinkerhoff and J.M. Brinkerhoff, "Public Sector Management Reform in Developing Countries: Perspectives Beyond NPM Orthodoxy", Public Administration and Development. vol. 34, pp. 222–237. October 2015, doi: 10.1002/pad.1739.

[34] G. de Vries, M.Timmer, and K. de Vries, "Structural Transformation in Africa: Static Gains, Dynamic Losses", The Journal of Development Studies, vol.51, pp.674-688. 2015, doi: 10.1080/00220388.2014.997222

[35] L.A. Ika and J. Donnelly, "Success conditions for international development capacity building projects", International Journal of Project Management", vol. 35, pp. 44-63. January 2017, doi.org/10.1016/j.ijproman.2016.10.005

[36] R. Crook, "Rethinking civil service reform in Africa: 'islands of effectiveness' and organizational commitment", Commonwealth & Comparative Politics, vol. 48, pp. 479-504. 2010, doi: 10.1080/14662043.2010.522037

[37] F. Abuzeid, "Foreign aid and the 'big push theory': lessons from sub-Saharan Africa", Stanford Journal of International Relations, vol. 10, 2009, pp. 1–9.

[38] T. Hopper, "Neopatrimonialism, good governance, corruption and accounting in Africa", Journal of Accounting in Emerging Economies, vol. 7, pp. 225-248. May 2017, https://doi.org/10.1108/JAEE-12-2015-0086

[39] P. Yanguas and B. Bukenya, "New' approaches confront 'old' challenges in African public sector reform", Third World Quarterly, vol. 37, pp. 136-152. 2016, doi: 10.1080/01436597.2015.1086635

[40] V. Hasheela-Mufeti, "Current prospects and challenges of enterprise resource planning (ERP) adoption in developing countries", International Science & Technology Journal of Namibia, vol. 9, pp. 94-106. 2017, doi: http://hdl.handle.net/11070/2135

[41] A. Eberhard, K. Gratwick, E. Morella, and P. Antmann, "Independent power projects in Sub-Saharan Africa: Investment trends and policy lessons", Energy Policy, vol. 108, pp. 390–424. September 2017, doi: 10.1016/j.enpol.2017.05.023

[42] H. Rosling, A.R. Rönnlund, and O. Rosling, Factfulness: "Ten reasons we're wrong about the world-and why things are better than you think", New-York: Flatiron Books, 2018.

[43] A. Eberhard, K. Gratwick, E. Morella, and P. Antmann, "Independent power projects in Sub-Saharan Africa: Investment trends and policy lessons", Energy Policy, vol. 108, pp. 390–424. September 2017, doi: 10.1016/j.enpol.2017.05.023

[44] A. L. Zouaghi, "Aligning Key Success Factors to ERP Implementation Strategy: Learning from Case Study", International Journal of Business Information Systems vol. 22, pp. 100-115. January 2016, doi: 10.1504/IJBIS.2016.0757.

Appendix

Table I: Aggregated results of CSF of ERP implementation in African PSO

Dimensions	Sub-dimension (if applicable)	Critical Success Factor
A. SOFTWARE		1. Participatory software development/Testing/Troubleshooting 2. Fair and balanced ERP vendors/suppliers relationships 3. Country-related functional requirements 4. Adequate ERP infrastructure/hosting 5. Sufficient IT organizational maturity
B. PROCESS FLOW	**B1 Process**	1. Fit between ERP and an organization's procedures 2. Harmonized practices/procedures/processes 3. Good communication management processes
	B2 Data	4. Efficient data quality control 5. Good data collection processes and methods 6. Solid data management practices 7. Clear data conversion plan and management
C CUSTOMER MINDSET	**C1 User influence**	1. Users' attitudes/Openness to change; 2. Adequate technical competencies and knowledge of users 3. Access to training
	C2. Team influence	4. Adequate team member competencies 5. Team composition 6. Stability of teams/Low attrition rate 7. Good collaboration 8. Leadership
	C3. Organizational influence	9. Prior experience in ERP/major IT project implementation 10. Change management competency 11. Organizational support/commitment 12. Presence of a champion 13. Shared vision/ mission/ organizational goals 14. Stakeholder's ownership of the project 15. Need driven endeavor;
D CHANGE MANAGEMENT	**D1. User attitude**	1. Effective communication the change to users 2. Adequate training and education of users 3. Active participation of users
	D2. Project (scope) changes	4. Effective change control management processes and procedures
	D3. Business process changes	5. Harmonization of practices and processes 6. Assessment of best practices
	D4. System changes	7. Management of interests 8. Communicate change throughout the organization 9. Plan and manage corporate culture change
E EXTERNAL ENVIRONMENT		1. Fit with national culture and values 2. Balanced donor-recipient relations 3. Adequate local infrastructure
F METHODOLOGY		1. Good project management 2. Clear implementation strategy

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601