MAPPING HEALTH STATISTICS: REPRESENTING DATA RELIABILITY

Alan M. MacEachren
Cynthia A. Brewer
Department of Geography
Penn State University
University Park, PA 16802
USA
[MacEachren e-mail: nyb@psuvm.psu.edu]
[Brewer e-mail: cbrewer@essc.psu.edu]

Linda W. Pickle
National Center for Health Statistics, CDC
Hyattsville, MD
USA
[Pickle e-mail: LWPØ@NCHØ9A.EM.CDC.GOV]

Abstract

Data reliability is a major concern for both science and policy analysis. Methods of specifying the reliability of sample data, the variance around measures of central tendency, the confidence we should put in statistical summaries, etc. are well developed. When data are geo-referenced, however, reliability estimates have not traditionally been mapped. For the same reasons that we map spatial data rather than limiting ourselves to tables or to numerical results of statistical analysis, we should portray data reliability in map form. This paper reports on the first stage of an effort to develop and assess reliability representation methods in the context of the U. S. National Center for Health Statistics *Mortality Atlas*.

1 Introduction

Reliability representation (under a variety of labels, including data quality visualization and representation of uncertainty) is a topic that has received increasing attention from researchers over the last several years,[1] but one for which no clear guidelines have yet been established. Research conducted thus far has focused on two separate (but related) issues, what to represent and how to represent it.

1.1 Reliability of geo-referenced information: what to represent

Reliability of geo-referenced information has several components, each of which may require different representation methods. Several attempts have been made to delineate these components. The best known is probably that incorporated in the U. S. Spatial Data Transfer Standard (SDTS), in which emphasis is placed on data "quality" [8]. The SDTS identifies five components of data quality: positional accuracy, attribute accuracy, logical consistency, completeness, and lineage.

[1] See for example, papers by Buttenfield and Beard [4] for a proposed research agenda, MacEachren [11] and McGranaghan [18] on representation methods, Fisher (1994) on sonic symbols to embed information about data quality in maps, and van der Wel et al. [22] on syntactics for linking graphic variables to kinds of data reliability.

The focus on data quality, as it is defined in the SDTS, leaves out several important components of reliability that are likely to be critical to map users. Among these are the temporal aspect of data and their spatial, temporal, and attribute *resolution* [11]. A single value representing a census block, for example, is a more reliable summary for that block's inhabitants than is a single number representing an entire state. In addition to including time and resolution issues, data reliability (or quality) can only be assessed against a specific "data model." Buttenfield and Beard [5] propose three categories of data models relevant for geo-referenced information, *continuous, multinomial (categorical), and discrete*. In the context of data for contiguous enumeration units, MacEachren and DiBiase [13] suggested a two-dimensional phenomenon space defining a range of data models on the basis of spatial continuity (a continuous–discrete dimension) and characteristics of variation across space (an abrupt–smooth dimension). MacEachren [11] describes the potential application of this data model framework to categorizing the spatial aspects of data reliability.

1.2 Reliability of geo-referenced information: how to represent

Once we select the components of reliability information to represent and determine how to assess reliability within these components (the what of reliability representation) we must deal with how to represent. This step involves two interrelated decisions. The first relates to the way data and reliability information (metadata) are linked and accessed – the interface style. The second deals with the way the metadata are signified – the symbolization style.[2]

The "interface" decision begins with a choice about whether data and metadata will occupy the same display location or separate locations. If data and metadata are separated, there is a further choice between representing the metadata graphically (as a second map) or nongraphically (in verbal or tabular form, indexed to map locations). Data and metadata can occupy the same location through visual *overlay* or *merger*. The result can be considered a form of bivariate map. In a bivariate overlay, the data and metadata are signified with distinct symbol sets that remain visually *separable* when one is superimposed upon the other (e.g., a coarse line or dot pattern superimposed on an area fill). Merger involves creating a symbol *conjunction* in which unique symbols depict each possible combination of data and reliability. For a seven-class map with two levels of data reliability, the legend would show 14 categories.

The above discussion applies to both paper and electronic maps. For electronic maps, however, the interface issues expand along with the greater flexibility provided. One added issue relates to when metadata is available and whether its appearance/disappearance is under user control. Through interaction, for example, metadata can be linked to the data depiction but remain invisible unless accessed using a probe that responds with a symbol, sound, or value as a user points to a data location. This method of information access can be extended to probes that provide continuous feedback while they are moved across a map (e.g., Fisher's, [9] sonic

Metadata are "data about data." Measures of data reliability, therefore, are one form of metadata. Below, the term "metadata" will be used interchangeably with "reliability information." In most cases, however, the representation issues discussed for depicting reliability apply to all kinds of metadata representation, whether they deal with data reliability or some other attribute of the data.

probe that allows an analyst to "play" the reliability information underlying a classified satellite scene by dragging a probe across the image). Interaction also makes it possible to toggle a full metadata depiction on and off, either switching between separate data and metadata layers of a bivariate overlay or turning the metadata component of a merged map on and off [14]. Interactive controls can also allow an analyst to manipulate a reliability threshold, thus using the Exploratory Data Analysis (EDA) technique of "focusing" to visually target specific reliability levels. Beyond direct interaction, dynamic display provides for the possibility of animated sequential presentation of data and metadata (or alternative views of data) as a way to represent the spatial distribution of pattern "stability" [7].

In addition to these interface issues, decisions must be made concerning the specific symbolization methods for depicting reliability. Clearly, these decisions are integrally linked to the symbolization choices made for the data (e.g., data represented on a choropleth map produced in full color will require different reliability representation methods than data depicted on a dot map produced in black and white). Several authors have addressed the question of symbol methods for reliability representation at a conceptual level. MacEachren [11], McGranaghan [18], and van der Wel et al. [22] have all presented frameworks based on Bertin's [1] graphic variable typology (with some extensions). MacEachren argued that two graphic variables missing from Bertin's original set, color saturation and clarity,[3] are particularly well suited to depicting data reliability – because both have the potential to suggest uncertainty or lack of precise knowledge. A comprehensive syntactics for reliability representation is provided by van der Wel et al. [22]. This syntactics specifies appropriate links between graphic variables and kinds of reliability information.[4]

2 The U. S. National Center for Health Statistics – Mortality Atlas

Our research extends the general approach to reliability representation outlined above to the specific context of paper maps that depict mortality statistics. Before discussing the specifics of the representation forms being evaluated, therefore, a brief introduction to this map use context is in order.

Death rate maps have been demonstrated to be useful visualization prompts.[5] These maps allow analysts to identify cancer "hot spots" and spatial patterns that, in many cases, had not been apparent in tabular mortality statistics. These visually identified hot spots and patterns

[3] The term "clarity" is used here rather than "focus" (the term originally used in the paper cited). Focus was found to be a problematic term on several counts, among them was potential confusion with the EDA term "focusing." See MacEachren [12] for further explanation.

[4] The term "syntactics" as used here refers to a conceptual framework for relating attributes of and relations among sign-vehicles (map symbols) to corresponding attributes of and relations among referents (the things signified by the map symbols).

[5] Most published evidence of the role of mortality rate maps in health research is related to use of traditional paper maps (rather than to interactive electronic maps). Some examples are the series of cancer death rate maps published by the U. S. National Cancer Institute [16, 17, 19, 20] and those published by the Chinese [6].

become hypotheses that can be tested through case-control studies designed to determine whether apparent patterns are real and, if so, the reasons for high rates in particular regions [10]. This combination of map and statistical analysis has led to such important discoveries as the association between oral cancer and snuff dipping, the association between lung cancer and exposure to asbestos through shipyard work in the U. S. during World War II, and the link between dietary factors and esophageal cancer in the Chinese province of Linxian. The success of the early cancer atlases led to the publication of similar atlases in most developed countries of the world [23].

The demonstrated usefulness of mortality atlases is the impetus behind the project discussed here, a U. S. atlas of leading causes of death. The atlas is being prepared by the National Center for Health Statistics (NCHS), the U. S. federal agency responsible for collecting and publishing information from all U.S. death certificates. Due to differences in population age distributions of areas to be mapped, death rates depicted in the atlas require some age-adjustment. Directly-adjusted rates (the result of multiplication of each area's age-specific rates by a common standard population) were selected for mapping because they are comparable across causes. The new atlas will include maps of approximately 20 leading causes of death, as they have been defined for previously-published NCHS mortality statistics. Data for 1988-1992 will be included by race and gender. Data mapped are aggregated to Health Service Areas (HSA) – 805 groupings of counties based on a recent cluster analysis of where residents obtained hospital care in 1988 [15]. Clusters of counties were defined so that residents of an HSA were more likely to seek hospital care within that HSA than outside it. Three map types will be used in the atlas, full page (8.5 by 11 inches) seven-category choropleth maps depicting the directly-adjusted death rates, quarter page five-category choropleth maps showing results of significance tests comparing each HSA rate to the overall U. S. rate, and smoothed maps designed to highlight broad regional patterns.

Before committing funds to production of this mortality atlas, a conceptual framework for selecting specific hue combinations for diverging and sequential color schemes was required.[6] The framework developed places emphasis on avoiding potential perceptual and conceptual problems that might be encountered by the range of readers for whom these maps are being produced [21, 3]. Our goal in the research described here is to build from this base to devise effective reliability representation that is compatible with the symbolization, format, and color schemes to be used for the non-smoothed choropleth depictions in the atlas.

3 Representation Methods for Mortality Map Reliability

There are a variety of initial constraints that we needed to consider in developing potential reliability representation schemes for the planned atlas. Among the most important were that

[6] The terms sequential and diverging are taken from Brewer's [2] color use guidelines (a syntactics of logical matches between color schemes and kinds of mapped information). Sequential schemes are those in which the goal is to emphasize the rank order of a data set through a sequence of colors visually ordered from high to low. Diverging schemes are those in which the goal is to emphasize progression outward from a critical midpoint in a data range.

the atlas will be a paper product, printed in color, and will use choropleth symbolization to depict the death rates. In relation to the issue of "what to represent," then, the assumed data model is one having a continuous but abruptly varying distribution. The component of reliability of interest in relation to this data model is the accuracy of values aggregated to HSAs as a representation of death rates for the entire HSA. Thus, a measure of variance around these single values for each HSA is the appropriate reliability index for which a representation method is required. Figure 1 illustrates the range of reliability present in a subset of the available data (based on a coefficient of variation threshold of 20 percent). Clearly, reliability of these maps varies spatially and differs substantially from variable to variable.

Emphasis in this project is on developing representation schemes in which reliability information can be embedded in (or overlaid upon) a choropleth map of the data. Such schemes can be considered a special case of bivariate mapping in which two characteristics of one variable are mapped (e.g., mean and variance). Several bivariate data–reliability schemes are being assessed against a depiction of data and reliability individually on map pairs (color value, either diverging or sequential, for death rates and a separate two-class gray tone map for unreliable death rates).

Cognitive theories of *selective* and *divided* attention (concepts related to Bertin's notions of *selectivity* and *associativity*) suggest hypotheses about the ability of map users to extract information from specific kinds of bivariate representations. Some combinations of representation methods should make it easy to focus on data or metadata independently (e.g., color value for data and texture for metadata), while others should facilitate integration of the data–metadata information (e.g., color hue plus color saturation). In addition to assessing the general usefulness of different representation schemes, therefore, the experiment being conducted provides an opportunity to evaluate theoretical predictions (based on theories of selective and divided attention) in the context of a practical cartographic application.

We began by considering eight potential data–reliability representation schemes for bivariate data/reliability maps (to be compared with data/reliability map pairs). These initial schemes are listed below.[7]

1. distinct color hues for death rates, higher color value of the same hue for unreliable death rates
2. color value (either diverging or sequential) for death rates, desaturated (grayed) versions of the same values of hues for unreliable death rates

[7] The first scheme listed was used by Pickle, et al. [21] to demonstrate the potential for merging data and reliability information on the same map of health statistics. The remaining schemes are all based on use of color value as the primary variable to represent death rate categories. Color value alone results in a gray scale, a sequential scheme. Adding a hue difference, however, can enhance the discriminability of a sequential scheme that differs primarily in value (e.g., a range from light yellow through orange to dark red). Diverging schemes, although also relying primarily on color value differences, require a hue difference so that users can distinguish between values above and below the midpoint. The most effective diverging schemes are modeled on pairs of sequential schemes joined by a light neutral color at the midpoint (e.g., dark red, medium red, light red, light gray, light blue, medium blue, dark blue).

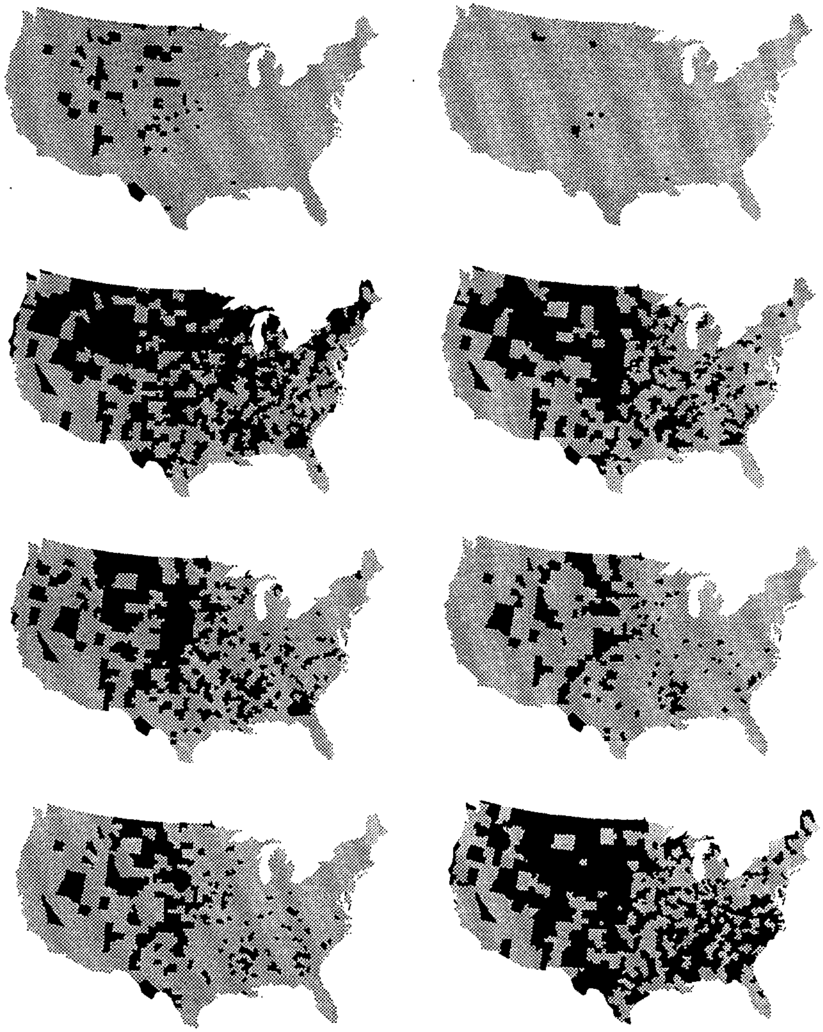

Figure 1. These maps depict reliability of death rates aggregated to Heath Statistical Areas for eight sample causes of death. In each case, areas shaded black are deemed unreliable (in relation to a coefficient of variation threshold set at 20%).

3. color value (either diverging or sequential) for death rates, gray tones (complete desaturation) of the same color value for unreliable death rates
4. color value (either diverging or sequential) for death rates, hue shift for unreliable death rates (e.g., blue for reliable data shifted to purple for unreliable)
5. color value (either diverging or sequential) for death rates, texture overlay (in black) for unreliable death rates
6. color value (either diverging or sequential) for death rates, texture "embedded" within shading (in white) for unreliable death rates
7. color value (either diverging or sequential) for death rates (of sufficient texture for individual dots to be visible), hue underlay (in a distinctive hue such as yellow) for unreliable death rates
8. color value (either diverging or sequential) for death rates, single point symbols (e.g., asterisks) in HSAs for unreliable death rates

Not all of the data/reliability representation schemes initially considered proved to be suitable choices as additional parameters of the planned atlas were determined. These parameters include: a need for the reliability representation method to work with both sequential and diverging color schemes for the data, production of discriminable area fills in HSAs on quarter page maps, and practicality for printing paper maps using offset lithographic printing (in four inks).

In relation to the parameters identified, we reduced the initial set of eight data/reliability schemes to three that are now being empirically assessed and compared to map pairs depicting data and reliability separately. These schemes are those listed above as numbers 2 (value/saturation), 5 (value/texture overlay), and 4 (value/hue shift).

5 Next Steps

Reliability representation is a critical research area as an increasing variety of users are relying on maps as tools for prompting hypotheses and formulating policy. Although the research discussed here is targeted at a particular kind of map and map use, we anticipate that results will be useful in related reliability representation contexts.

The initial stage of the project outlined here focused on delineating a framework for making logical choices among representation methods and selecting four potentially effective data/reliability representation schemes suited to the choropleth maps planned for the NCHS *Mortality Atlas*. The second stage of the research involves an empirical assessment of the four scheme types arrived at above for use on choropleth maps generated for the Health Service Areas in the conterminous U. S. This portion of the research will assess the data/reliability representation methods through comparison of map interpretation performance on typical map reading tasks. These tasks represent three levels of task complexity (rate retrieval for individual map areas, region comparison within a map, and comparison of different maps). More specifically, tasks are designed to answer the following questions:

- does having reliability information influence rate retrieval for the data
- can users determine which map category is least/most reliable overall
- can users estimate reliability of a region (e.g., judge the percent reliable)
- does presence of reliability information influence region comparison tasks directed to the data
- does the way in which reliability information is depicted influence region comparison tasks directed to the data
- does having reliability information influence overall map pattern assessment tasks directed to the data (e.g., about map clustering or complexity)
- does providing reliability information decrease judgements about "truth" of the map as a whole? For example, are subjects less likely to judge a map as 'reliable enough' to make decisions about allocation of health care funds if reliability information is present? Or, is judgement of "truth" linearly related to the proportion of health units having a coefficeint of variation of 20 percent or higher.

In addition to the above tasks, preferences for the different data/reliability representations schemes will be determined. Attention will be directed to preferences for particular reliability representation methods and to the potential that adding reliability representation may influence preferences for color schemes applied to data representation.

Results of the empirical analysis will be presented in Barcelona.

References

[1] Bertin, J. 1983. *Semiology of Graphics: Diagrams, Networks, Maps.* Madison, WI: University of Wisconsin Press.

[2] Brewer, C. A. 1994. Color use guidelines for mapping and visualization. In *Visualization in Modern Cartography,* ed. A. M. MacEachren and D. R. F. Taylor. London: Pergamon, pp. 123-147.

[3] Brewer, C. A., MacEachren, A. M. & Pickle, L. W. (1995). Evaluation of map color schemes for the NCHS mortality atlas. In *Proceedings of the International Symposium on Computer Mapping in Epidemiology and Environmental Health.*

[4] Buttenfield, B. and Beard, M. K. 1991. Visualizing the quality of spatial information. *Auto-Carto 10* Baltimore, Maryland, pp. 423-427.

[5] Buttenfield, B. P. and Beard, M. K. 1994. Graphical and geographical components of data quality. In *Visualization in Geographic Information Systems,* ed. D. Unwin and H. Hearnshaw. London: John Wiley & Sons, pp. 150-157..

[6] Editorial Committee for the Atlas of Cancer Mortality. 1979. *Atlas of Cancer Mortality in the People's Republic of China,* Shanghai, China: China Map Press.

[7] Evans, Beverley 1995. Cartographic display of data certainty: Does it benefit the map user? paper presented at the Annual Meeting of the Association of American Geographers, 14-18 March, 1995, Chicago, Illinois.

[8] Fegeas, R. G., Cascio, J. L. and Lazar, R. A. 1992. An overview of FIPS 173, The spatial data transfer standard. *Cartography and Geographic Information Systems* 19(5): 278-293.

[9] Fisher, P. 1994. Hearing the reliability in classified remotely sensed images. *Cartography and Geographic Information Systems* 21(1): 31-36.

[10] Hoover, R. Mason, T. J., McKay, F. W., Fraumeni, J. F. Jr. 1975. Cancer by county: New resource for etiologic clues. Science 189(4207): 1005-1007.

[11] MacEachren, A. M. 1992. Visualizing uncertain information. *Cartographic Perspectives* (13): 10-19.

[12] MacEachren, A. M. 1995. *How Maps Work: Issues in Representation and Design*. New York: Guilford Press.

[13] MacEachren, A. M. and DiBiase, D. W. 1991. Animated maps of aggregate data: Conceptual and practical problems. *Cartography and Geographic Information Systems* 18(4): 221-229.

[14] MacEachren, A. M., Howard, D., von Wyss, M., Askov, D. and Taormino, T. 1993. Visualizing the health of Chesapeake Bay: An uncertain endeavor. *Proceedings, GIS/LIS* '93 Minneapolis, MN, 2-4 Nov., 1993, pp. 449-458.

[15] Makuc, D., Haglund, B., Ingram, D. D., Kleinman, J. C., & Feldman, J. J. (1991). *Health Service Areas for the United States*. National Center for Health Statistics. Vital and Health Statistics, Series 2, No. 112 (DHHS Publication No. 91-1386). Washington, DC: U. S. Government Printing Office.

[16] Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J., Fraumeni, J. F., Jr. (1975). *Atlas of Cancer Mortality for U.S. Counties: 1950-1969*. DHEW Publ. No. (NIH) 75-780. Washington, D.C.: U.S. Govt. Printing Office.

[17] Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J., Fraumeni, J. F., Jr. (1976). *Atlas of Cancer Mortality Among U.S. Nonwhites: 1950-1969*. DHEW Publ. No. (NIH) 76- 1204. Washington, D.C.: U.S. Govt. Printing Office.

[18] McGranaghan, M. 1993. A cartographic view of spatial data quality. *Cartographica* 30(2 & 3): 8-19.

[19] Pickle, L.W., Mason, T.J., Howard, N., Hoover, R., Fraumeni, J.F., Jr. (1987). *Atlas of U.S. Cancer Mortality among Whites, 1950-1980.* DHHS Publ. No. (NIH) 87- 2900. Washington, D.C.: US Government Printing Office.

[20] Pickle, L.W., Mason, T.J., Howard, N., Hoover, R., Fraumeni, J.F., Jr. (1990) *Atlas of U.S. Cancer Mortality among Nonwhites, 1950-1980.* DHHS Publ. No. (NIH) 90- 1582. Washington, D.C.: US Government Printing Office, 1990.

[21] Pickle, L. W., Herrmann, D., Mungiole, M., White, A. (1994) Design of the New U.S. Mortality Atlas. In *Proceedings of the International Symposium on Computer Mapping in Epidemiology and Environmental Health.*

[22] van der Wel, F. J. M., Hootsman and Ormeling, F. 1994. Visualization of data quality. In *Visualization in Modern Cartography*, ed. A. M. MacEachren and D. R. F. Taylor, London: Pergamon, pp. 313-331.

[23] Walter, S. D. and Birnie, S. E. 1991. Mapping mortality and morbidity patterns: An international comparison. *International Journal of Epidemiology*, 20: 678-689.