[AM-09-107] Spatial Data Uncertainty

Abstract

Although spatial data users may not be aware of the inherent uncertainty in all the datasets they use, it is critical to evaluate data quality in order to understand the validity and limitations of any conclusions based on spatial data. Spatial data uncertainty is inevitable as all representations of the real world are imperfect. This topic presents the importance of understanding spatial data uncertainty and discusses major methods and models to communicate, represent, and quantify positional and attribute uncertainty in spatial data, including both analytical and simulation approaches. Geo-semantic uncertainty that involves vague geographic concepts and classes is also addressed from the perspectives of fuzzy-set approaches and cognitive experiments. Potential methods that can be implemented to assess the quality of large volumes of crowd-sourced geographic data are also discussed. Finally, this topic ends with future directions to further research on spatial data quality and uncertainty.

Keywords: accuracy, error, geo-semantic uncertainty, quality, spatial data quality, spatial data uncertainty

Author & citation

Li, L. (2017). Spatial data uncertainty. The Geographic Information Science & Technology Body of Knowledge (4th Quarter 2017 Edition), John P. Wilson (ed). DOI: 10.22224/gistbok/2017.4.4

Explanation

1. Definitions
2. Background
3. Uncertainty modeling, simulation, and visualization
4. Geo-semantic uncertainty
5. Uncertainty in big geospatial data
6. Future directions

1. Definitions

spatial data quality: Evaluation of the similarity between spatial data and geographical truth, including both positional truth and attribute truth. The closer spatial data is to the truth, the higher its quality.

spatial data uncertainty: Uncertainty implies that there is something we are not sure of in spatial data and analysis due to various reasons, such as ignorance of human knowledge, generalization of geographic features, measurement errors, and incomplete representation of all factors in analysis.

spatial data accuracy: A measurement of error and usually used to communicate positional uncertainty. For example, a dataset with 10m accuracy means the represented

location of a feature may be as far as 10m from its true location.

spatial data error: A term closely related to spatial data accuracy. The smaller the error, the higher the accuracy, and the better the data quality.

geo-semantic uncertainty: Uncertainties associated with geographic concepts, classes, and values, such as vague boundaries of “downtown.”

2. Background

People tend to believe that computer outputs are reliable, at least more reliable than human interpretations. When a map is created from GIS software or results generated from a spatial model, the evidence seems so compelling and objective. Most of the time, we rely on datasets collected and generated by other agencies and never question data quality or evaluate the appropriateness of a particular dataset for tasks at hand. However, from conceptualization to generalization, from measurement to analysis, information loss is unavoidable; therefore, all representations of the world are imperfect. As demonstrated in Figure 1, apparent positional discrepancies may be present in street network datasets produced by two different agencies. It is critical to identify, assess, and quantify uncertainty in spatial data and analysis because not accounting for uncertainty can lead to overconfidence in the conclusions.

Figure 1. Positional uncertainty: Street networks in Santa Barbara, CA from two different data sources (Li & Goodchild, 2011).

In the GIScience community, the issue of spatial data uncertainty has increasingly attracted researchers over the past few decades. The first conference dedicated to address this area probably dates back to the 1980s when “Accuracy of Spatial Databases” was organized by the U.S. National Center for Geographic Information and Analysis (NCGIA) in Santa Barbara (Goodchild & Gopal, 1989). In addition, twelve International Symposiums on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences and nine International Symposiums on Spatial Data Quality have been organized biannually. Another major contributor on spatial data uncertainty has been the Conference on Spatial Information Theory (COSIT) that was established in 1993, especially on the role of uncertainty and vagueness in geo-semantics. Uncertainty in geographic information was, and still is identified as one of the long-term research challenges in the University Consortium for Geographic Information Science (UCGIS) research agenda (McMaster & Usery, 2004).

3. Uncertainty modeling, simulation, and visualization

A range of approaches have been developed to study both positional and attribute uncertainty. Positional uncertainty has been evaluated using analytical approaches and simulations. Analytical methods use mathematic formulae and statistical principles to describe systematic and random errors. For vector datasets that include point, polyline, and polygon, all analytical approaches are based on error representation of coordinate pairs. Error ellipse model and its derivatives have been developed to represent uncertain positions of a point using the variances in the x and y directions and the covariance of x and y coordinates (Alesheikh et al., 1999), with the center of the ellipse being the most likely position of a point. To represent uncertainty of polylines, different models have been developed based on point uncertainty models, including the epsilon-band model, the error band model, and the G-band model (Perkal, 1966; Dutton, 1992; Shi & Liu, 2000). The general idea of polyline uncertainty is a buffer around linear features, which is called the epsilon band. The buffer size of an epsilon band is dependent on factors including positional uncertainty of each node and spatial autocorrelation of points along a line. Positional uncertainty of polygons is usually caused by uncertainties of nodes and lines that compose a polygon, so it can be assessed using similar methods mentioned above. Furthermore, methods have been developed to evaluate characteristics of polygons, such as area, perimeter, and centroid. For instance, uncertainty of polygon area has been studied using the variances of its vertices and covariance between them (Chrisman & Yandell, 1988; Zhang & Kirby, 2000).

Analytical methods for representing positional uncertainty is rigorous and supported by statistical theories; however, the assumption of independent uncertainties of all vertices is rarely the case. Moreover, mathematical formulae can quickly become too complex to implement with complicated feature shapes and may not be the best way to communicate uncertainty with general data users. Another way to represent positional uncertainty is through simulation that generates a large number of realizations of geographic positions based on random variables. The most widely adopted simulation method is Monte Carlo simulation that involves several steps (Alesheikh, 1998): find the probability density function of errors in the input data, generate a set of random variables drawn from the probability density function, perform spatial operations to generate N output realizations of the random variables, and calculate summery statistics from the N realizations. For

example, Monte Carlo simulation can be used to evaluate propagation of uncertainties of slope and aspect that are calculated from digital elevation models (DEMs).

An error matrix, also known as a confusion matrix, is a common way to evaluate uncertainty of categorical attributes. It is a matrix of represented categories versus true categories for a sample of geographic features, which may be objects or pixels in raster datasets. For instance, each row records every represented category in the dataset while each column records every true category. The position of every sampled data point in this matrix indicates its assigned category and true category. Error of commission (incorrect inclusion of data points in a true category) and error of omission (incorrect exclusion of data points in a true category) can be easily calculated from an error matrix. These two statistics tell us the accuracy of categorical attributes.

Visual representation is also important for describing spatial data uncertainty. MacEachren (1992) recommended four graphics variables to convey spatial uncertainty: contour crispness, fill clarity, fog, and resolution. Further, MacEachren et al. (2005) explored various cartographic tools to effectively communicate both geographic data and uncertainty simultaneously, particularly from the perspectives of conceptualization and decision-making. More recently, MacEachren et al. (2012) evaluated their use of visual semiotics to characterize different types of uncertainty based on two empirical studies.

4. Geo-semantic uncertainty

Geo-semantic uncertainty generally refers to ambiguities and vagueness associated with geographic concepts, classes, and values that are sometimes subjective without agreed definitions. Examples of such concepts include hill and mountain (what elevation threshold distinguishes a hill from a mountain), downtown (where is the boundary of downtown Long Beach), and exurbanization (more than eighteen definitions are available according to Berube et al., 2006). It is not straightforward to measure this type of uncertainty using statistics or mathematic formulae. Geo-semantic uncertainty is usually studied using fuzzy-set approaches and cognitive experiments. Fuzzy sets allow an element to partially belong to a set with the aid of a membership function ranging from zero to one. For example, a set of membership functions of distances can be constructed to represent and visualize the ambiguities associated with the concept of exurbanization (Ban & Ahlqvist, 2009). As a result, various degrees of grayness may be used to represent the likelihood of a particular location being exurbanized or not exurbanized, which cannot be effectively visualized as crisp boundaries (Figure 2).

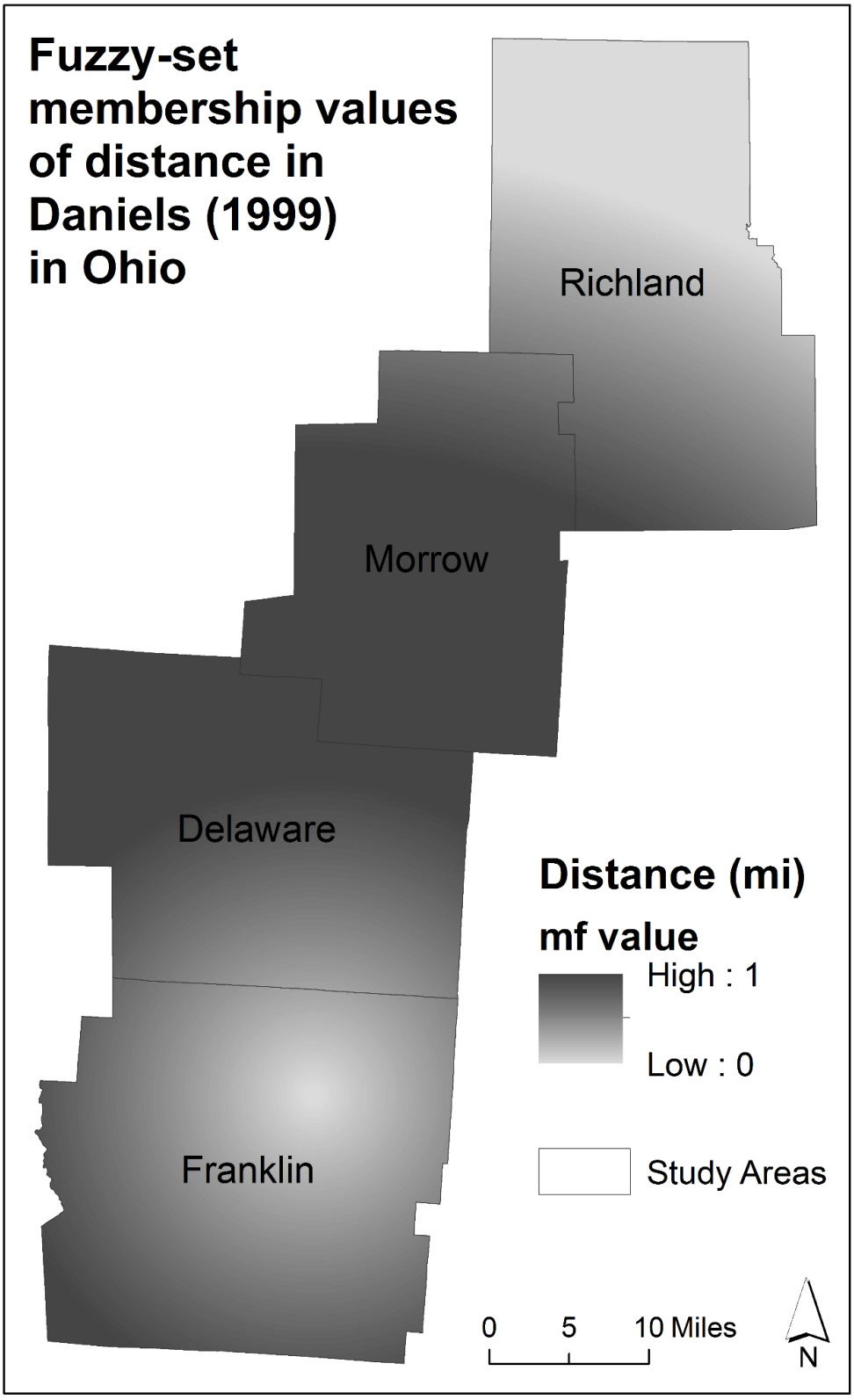

Figure 2. Uncertain boundaries of exurbanization using a fuzzy-set membership function. Darker color indicates higher degree of exurbanization (Li et al., 2018).

In addition to fuzzy-set approaches to study vague geographic concepts, behavioral and cognitive methods have also been adopted to visualize and evaluate geo-semantic uncertainty. This line of research relies on surveys and interviews to empirically determine the content of a vague spatial concept such as the extent of a region or spatial relations such as far and near. For example, Montello et al. (2003) recruited 36 participants to draw

the outline of downtown Santa Barbara on a base map and aggregated these maps to visualize vague boundaries using dot-density shading. A more recent study by him and his colleagues examined the vague regions of “Northern” and “Southern” California using a similar approach (Montello et al., 2014). Meanwhile, there is a trend to use crowd-sourced geographic data to complement formal questionnaires and interviews to elicit people’s perception of vague spatial concepts. Li and Goodchild (2012) identified vague boundaries of several major places in France using geotagged photos collected from Flickr. Gao et al. (2017) replicated the study of northern and southern California regions using data from five sources of user-generated content and obtained similar results.

5. Uncertainty in big geospatial data

Uncertainty has always been a fundamental issue in GIScience. However, it is of special importance to assess uncertainty in big geospatial data, sometimes referred to as Volunteered Geographic Information (VGI, Goodchild, 2007), due to lack of quality control in data creation. Goodchild and Li (2012) proposed three methods to evaluate the quality of VGI. First, based on the crowd-sourcing approach, higher data accuracy is associated with a larger number of reviewers and editors. The effectiveness of this approach was confirmed by several studies on data quality of OpenStreetMap (e.g., Haklay, 2010). This method works best for areas or geographic features that attract interest from a large number of people, such as street networks in London. Second, the social approach establishes a hierarchy of data quality experts based on their performance and reputation in a particular data collection project. This approach mimics the system of social hierarchy in any profession, which arranges people in a relatively linear ladder. People at the top have more power and control over data management, especially in the case of a dispute. This approach has been adopted by many open and collaborative projects such as Wikipedia and OpenStreetMap. Although every user can create and edit any feature, people at a higher level make a decision regarding the quality of a particular piece of geographic information when a disagreement happens. Third, the geographic approach relies on geographic theories and principles to assess the quality of geographic data. A geographic fact inconsistent with geographic rules may be flagged as potentially inaccurate or incorrect. An example would be a bar mistakenly geotagged to a historical site location. Furthermore, open source geographic data may be compared with an authoritative or high-accuracy reference dataset to generate a quality control report, as demonstrated by Haklay (2010). One major limitation of this method is unavailability of a gold standard database as crowd-sourced geographic data cover a wide range of geographic themes not included in any official datasets.

6. Future directions

Uncertainty is undoubtedly an essential characteristic of any geographic datasets and affects spatial analysis and decision-making. Researchers and scientists have made many efforts to understand, identify, evaluate, quantify, and reduce spatial data uncertainty and to increase spatial data quality using both analytical and simulation approaches. Although significant advances have been made in this area since the 1980s, there are still many challenges to be overcome. Many models have been developed to either mathematically or

empirically assess positional accuracy while approaches for evaluating attribute uncertainty are still few, let alone models that assess both positional and attribute uncertainty coherently. Currently separate models are usually adopted to measure uncertainty in vector and raster datasets. In a real-world scenario, one application may require both vector and raster data in a single step. How to accurately assess uncertainty propagation that incorporates both data structures requires more work. In the era of big data, large amounts of geographic data may present a great opportunity to study physical environment and human society; however, new methods for understanding spatial data quality and uncertainty in crowd sources are to be developed and implemented. Rich geographic datasets are created on places without accurate or explicit positional information, novel computational and empirical approaches need to be discovered. Finally, although many models have been developed to investigate and communicate spatial data quality and uncertainty, they are not integrated into GIS software packages. Educating GIS professionals and geospatial technology users on the importance of spatial data quality and uncertainty still has a long way to go.

References

Alesheikh, A. A. (1998). Modeling and managing uncertainty in object-based geospatial information systems. Ph.D. Thesis. The University of Calgary, Alberta, Canada.

Alesheikh, A. A., Blais, J. A. R., Chapman, M. A., & Kariml, H. (1999). Rigorous Geospatial data uncertainty models for GISs. In Jaton, A. and Lowell, K. (Eds.), Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources (pp. 195-202). Chelsea Michigan: Ann Arbor Press.

Ban, H., & Ahlqvist, O. (2009). Representing and negotiating uncertain geospatial concepts–Where are the exurban areas? Computers, Environment and Urban Systems, 33(4), 233-246.

Berube, A., Singer, A., Wilson, J. H., & Frey, W. H. (2006). Finding exurbia: America's fast-growing communities at the metropolitan fringe. The Brookings Institution: Living Cities Census Series (October), 1-48.

Chrisman, N. R., & Yandell, B. S. (1988). Effects of point error on area calculations: A statistical model. Surveying and Mapping, 48, 241-246.

Dutton, G. (1992). Handling positional uncertainty in spatial databases. Paper presented at the 5th International Symposium on Spatial Data Handling, Charleston, S.C., USA.

Gao, S., Janowicz, K., Montello, D.R., Hu, Y., Yang, J-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., and Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science, 31:6, 1245-1271.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211-221.

Goodchild, M. F. & Gopal, S. (Eds.). (1989). The Accuracy of Spatial Databases. CRC Press.

Goodchild, M. F. & Li, L. (2012). Assuring the quality of volunteered geographic information. Spatial Statistics, 1, 110-120.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design, 37(4), 682–703.

Li, L., & Goodchild, M. F. (2011). An optimisation model for linear feature matching in geographical data conflation. International Journal of Image and Data Fusion, 2(4), 309-328.

Li, L., & Goodchild, M. F. (2012). Constructing places from spatial footprints. GEOCROWD '12: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, p. 15-21.

Li, L., Ban, H., Wechsler, S.P., & Xu, B. (2018). Spatial Data Uncertainty. In: Huang, B. (Ed.), Comprehensive Geographic Information Systems. Vol. 1, pp. 313–340. Oxford: Elsevier.

MacEachren, A. M. (1992). Visualizing Uncertain Information. Cartographic Perspectives, 13, 10-19.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. Cartography and Geographic Information Science, 32(3), 139-160.

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual Semiotics & Uncertainty Visualization: An Empirical Study. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2496-2505.

McMaster, R. B., & Usery, E. L. (Eds.). (2004). A Research Agenda for Geographic Information Science. 1st Edition. Boca Raton, Florida: CRC Press.

Montello, D. R., Friedman, A., & Phillips, D. W. (2014). Vague cognitive regions in geography and geographic information science. International Journal of Geographical Information Science, 28(9), 1802-1820.

Montello, D. R., Goodchild, M. F., Gottsegen, J., & Fohl, P. (2003). Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. Spatial Cognition & Computation, 3(2-3), 185-204.

Perkal, J. (1966). On the length of empirical curves. Paper presented at the Michigan Inter-University Community of Mathematical Geography, Ann Arbor, MI, USA.

Shi, W., & Liu, W. (2000). A stochastic process-based model for the positional error of line segments in GIS. International Journal of Geographical Information Science, 14(1), 51-66.