JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 20 (2020), pp. 97–102
doi:10.5311/JOSIS.2020.20.664

INVITED ARTICLE

How well do we really know the world? Uncertainty in GIScience

Michael F. Goodchild

Department of Geography, University of California, Santa Barbara, CA, USA

Received: March 2, 2020; accepted: April 9, 2020

Abstract: There are many reasons why geospatial data are not geography, but merely representations of it. Thus geospatial data will always leave their user uncertain about the true nature of the world. Over the past three decades uncertainty has become the focus of significant research in GIScience. This paper reviews the reasons for uncertainty, its various dimensions from measurement to modeling, visualization, and propagation. The later sections of the paper explore the implications of current trends, specifically data science, new data sources, and replicability, and the new questions these are posing for GIScience research in the coming years.

Keywords: uncertainty, accuracy, precision, spatial resolution, data science, synthesis

1 Introduction

As Alfred Korszybski put it [9], "the map is not the territory," or in today's digital era we might say that geospatial data are not geography, but merely a representation of it (for a recent, broad review of the significance of this simple comment in science see [13]). There are many reasons for this. For example, it is impossible to measure location perfectly without inheriting the errors of the measuring instrument, whether it be yesterday's sextant or today's GPS. Many of the data types on which GIScience relies—including data on land use, land cover, or soils—are not strictly replicable; if two experts were asked to make the same soil map independently they would not produce identical results. Geospatial data are scale-dependent and always subject to a combination of generalization, abstraction, or simplification. When Roger Tomlinson and IBM designed the original GIS, the Canada Geographic Information System, in the mid 1960s, they chose to regard the paper maps being input to the system as the truth [4], and made no allowances for uncertainties. Vector geospatial data inherit this assumption today, and the accuracy of a vector database is still

commonly assessed against the paper maps that were its source, rather than against the reality that they supposedly represent.

Early work on this topic emphasized accuracy [5], that is, the measurement of the differences between a representation and the truth, often termed error. This was broadened to uncertainty, however, when it became clear that many areas of GIScience, such as the soil maps discussed in the previous paragraph, lack a clearly defined concept of truth. The study of uncertainty should thus include vagueness, fuzziness, and related concepts. A distinction should be drawn between accuracy and precision: in this paper the term precision refers to the level of detail in the reporting of a measurement.

The problem of uncertainty in geospatial data has many dimensions. GIScience has inherited a set of longstanding practices and traditions in cartography, where much of the emotional satisfaction of maps may stem from their very lack of uncertainty, and their habit of presenting a cleaned-up, simplified world that shows every feature in its place. In turn, users of GIS and geospatial technologies in general are often reluctant to acknowledge uncertainty, perhaps expecting that results from a machine that operates to a precision of eight or sixteen decimal digits will be even more accurate than those obtained from analog maps. We see this every day in apps that report latitude and longitude to far more decimal places than is achievable even with the best measuring instruments, and pay no attention to the actual physical dimensions of the feature whose location is being reported.

While many fields of science are able to address uncertainties using simple statistical models, such as the normal distribution, spatial data present their own complications. Tobler's First Law of Geography asserts that "nearby things are more similar than distant things," but this same principle also applies to uncertainties. For example, digital elevation data are commonly subject to errors in the meter range, perhaps as high as ten meters. If these errors had an independent, distorting effect on every item of data, then the process of estimating slope by differencing adjacent elevations would be so subject to error as to be almost useless. But most errors are in reality strongly and positively autocorrelated, allowing estimates of slope to be made with reasonable confidence. The same behavior occurs for many types of spatial error, and is often expressed in terms of absolute and relative errors: relative errors over short distances tend to be much less than absolute errors. But while Tobler's statement is simple, the methods of dealing with autocorrelated errors are far from simple (see for example [14]).

These arguments point to the critical need for information about provenance in geospatial data. What measuring instruments were involved, and what were their average measurement errors? Which sources of uncertainty are likely to have become embedded in a dataset, and what patterns of autocorrelation have they produced? When two datasets are involved, have they been acquired independently or do they share some aspects of provenance (were they developed from the same base dataset, for example)? Significant strides have been made in the development and adoption of standards for geospatial metadata [2], but it remains difficult for users to assess whether one or more datasets are fit for an intended application.

Much research effort has gone into developing methods for visualizing the uncertainties present in geospatial data [11]. It is easy to imagine blurring features, or greying out uncertain attributes, but it is far more difficult to convey the essential property of positive spatial autocorrelation. Animation has proven a powerful option here. For example, Ehlschlager [1] was able to animate the uncertainty in the effects of sea-level rise on Boston

Harbor by showing a sequence of images, each of which represented a possible topography subject to known errors and spatial autocorrelations.

Finally, the models that increasingly underlie the development of policy are also subject to uncertainty of various kinds. There will be uncertainty in the input data, but also in the model itself in the form of missing variables or uncertain calibrations. Some insights can be gained by propagating data uncertainties [6] through the model, or evaluating the sensitivity of results to variation in data inputs. Another strategy is the ensemble approach, widely adopted in studies of climate change [8], which assumes that the variation in results across alternative and competing models somehow represents the distribution of result uncertainties.

In summary, uncertainty in geospatial data is a longstanding and much-researched problem with several important dimensions. It impacts all aspects of geospatial data and all applications, from everyday guidance apps to the modeling of global climate change. But despite several decades of progress, the impact on the broader user community and on the general public has been disappointing, and the community researching geospatial uncertainty remains small. There is little support for uncertainty in mainstream geospatial software products, though there is abundant support in the narrower fields of geostatistics and spatial statistics. As noted earlier, there may be good reasons for this, in continuing adherence to well-established cartographic practices, and in the complexity of dealing with spatially autocorrelated uncertainties. The remaining section of the paper looks forward, suggesting ways in which this situation might be improved in the coming years.

2 New directions and challenges

2.1 Uncertainty in data science

Data science has been growing very rapidly in recent years, with new programs and positions in universities, and a very strong demand for data scientists in industry. The relationship between GIScience and data science has been explored in numerous papers, and the phrase "spatial data science" is growing in popularity. The Fourth Paradigm [3, 7] argues for a new kind of science that is data driven, where theory emerges from analysis rather than driving it, and encourages scientists to "let the data speak for themselves." Artificial intelligence and deep learning are being promoted as ways of searching for patterns in data, and thus for making effective predictions.

Two major issues emerge from this argument for the GIScientist. First, we know that not all patterns are equally likely, and that the established principles of GIScience—spatial dependence, spatial heterogeneity, spatial resolution—limit what can be found in practice on the Earth's surface. What role should these principles play in the search for patterns? Take data on spatial interaction for example. Rather than using the standard spatial interaction model, one might generate a multitude of algebraic forms and test data against all of them, looking perhaps for the one that fits the data best. But many of these forms could be excluded a priori because of the required scaling behavior of spatial interaction models.

Uncertainty presents the second issue. How can scientific knowledge emerge in a field such as GIScience where all data are uncertain? How can techniques of artificial intelligence deal with the peculiar properties of spatial uncertainty? Should analysis begin with a collection of alternative data sets, each of which represents a possible true state of geographic reality that is consistent with the known uncertainties? Then how can the patterns

that emerge from each realization be compared and synthesized? And more broadly, what kinds of new geographic knowledge might emerge from the use of machine learning on geospatial data with explicit uncertainty?

2.2 New data sources

The old world of carefully documented, rigorously acquired data, exemplified by the census and by the work of national mapping agencies, is rapidly giving way to the world of Big Data, with its massive volumes and endless variety. Instead of a single source from which to answer a basic query, such as "what is the elevation of this point?", we now have access to digital elevation models from various sources at various resolutions, as well as digitized map contours and catalogs of spot heights. This leads to a new class of questions: how to integrate data from various sources, of various provenance and quality, into a single best estimate, and how to assess that estimate's uncertainty? Sui [12] argued that synthesis was becoming as important today as analysis may have been in the past; yet today the standard toolkit of spatial analysis still offers very few techniques that deal with data of varying quality. We all know how to estimate a mean and its standard deviation from data of uniform variance, but what is the best estimate of the mean and its standard deviation from data of non-uniform variances?

Our new data sources are often of dramatically improved spatial resolution, as for example when traditional origin-destination matrices are compared with the results of tracking individuals and vehicles. Much of the information being obtained from social media resolves to the individual. But while we might naively expect a corresponding reduction in uncertainty, in reality the new types of data introduce new types of uncertainty that are difficult to model using traditional techniques. For example, can GIScience develop models of the uncertainties present in the tracks of individuals? What kinds of models might be used to study uncertainty in the locations obtained from various implementations of GPS, or to interpolate between observations of the space-time location of an individual driver, passenger, or walker?

2.3 Replicability

The "replicability crisis" has recently generated very significant interest in science, for example in psychology [11]. Kedron et al. [8] have recently provided an extensive perspective on replicability in geographical analysis, and have made the concept of uncertainty a central feature of their discussion. While the paper focuses on geographical analysis, it is clear that many of the arguments resonate well in GIScience, and that GIScientists would do well to pay greater attention to replicability in how they design, execute, and report their work.

One of the central principles of GIScience is spatial heterogeneity [10], the observation that conditions vary across the surface of the Earth and that the results of analysis of one area do not necessary apply—are not necessarily replicable—in other areas. But if uncertainty is present in the data and therefore in the results of analysis, how much variation from one area to another is attributable to uncertainty, and how much to spatial heterogeneity? What is the role of place-based methods in this context, since they explicitly allow model parameters to vary from one area to another? Do we need a modified concept of replicability, call it weak replicability or weak generalizability, to accommodate the essential nature of research in GIScience?

3 Concluding remarks

The research of the past three decades has produced a rich body of knowledge regarding uncertainty in GIScience. But new developments are begging new questions, and it is clear that the growth of data science, the emergence of new data sources, and new concerns about replicability are stimulating a continued need for research. The second part of the paper outlined many of those new questions, and many more undoubtedly will emerge in the coming years as we continue to question the practices of the past, and to move beyond their legacy.

References

[1] EHLSCHLAEGER, C. R. *The stochastic simulation approach: Tools for representing spatial application uncertainty*. PhD thesis, University of California, Santa Barbara, 1999.

[2] FEDERAL GEOGRAPHIC DATA COMMITTEE. *Content standard for digital geospatial metadata workbook version 2.0. FGDC-STD-001-1998*. Federal Geographic Data Committee, 1998.

[3] GAHEGAN, M. Fourth paradigm GIScience? Prospects for automated discovery and explanation from data. *International Journal of Geographical Information Science 34*, 1 (2020), 1–21. doi:10.1080/13658816.2019.1652304.

[4] GOODCHILD, M. F. Reimagining the history of GIS. *Annals of GIS 24*, 1 (2018), 1–8. doi:10.1080/19475683.2018.1424737.

[5] GOODCHILD, M. F., AND GOPAL, S. *The accuracy of spatial databases*. Taylor & Francis, 1989.

[6] HEUVELINK, G. B. *Error propagation in environmental modelling with GIS*. CRC press, 1998.

[7] HEY, T., TANSLEY, S., TOLLE, K., ET AL. *The fourth paradigm: data-intensive scientific discovery*, vol. 1. Microsoft research Redmond, WA, 2009.

[8] KEDRON, P., FRAZIER, A. E., TRGOVAC, A. B., NELSON, T., AND FOTHERINGHAM, A. S. Reproducibility and replicability in geographical analysis. *Geographical Analysis* (2019). doi:10.1111/gean.12221.

[9] KORZYBSKI, A. *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. The International Non-Aristotelian Library Pub. Co, 1933.

[10] LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J., AND RHIND, D. W. *Geographic information science and systems, 4th edition*. John Wiley & Sons, 2015.

[11] MACEACHREN, A. M., ROBINSON, A., HOPPER, S., GARDNER, S., MURRAY, R., GAHEGAN, M., AND HETZLER, E. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science 32*, 3 (2005), 139–160. doi:10.1559/1523040054738936.

[12] SUI, D. Information synthesis. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*. Wiley Online Library, 2016, pp. 1–13.

[13] WUPPULURI, S., AND DORIA, F. A. *The Map and the Territory: Exploring the Foundations of Science, Thought and Reality*. Springer, 2018.

[14] ZHANG, J., AND GOODCHILD, M. F. *Uncertainty in geographical information*. Taylor & Francis, 2002.