GIS-assisted Data Analysis - Finding Meanings in Complex Spatial Data Sets

Tuija Kirkinen
Department of Archaeology, University of Helsinki.
P.O.Box 13, 00014 Helsinki, Finland.
E-mail: tuija.kirkinen@helsinki.fin

Introduction

Spatial Data in Archaeology is a sub-project of a large research consortium, *From Data to Knowledge* (1996-2001) funded by the Academy of Finland, and lead by professor Heikki Mannila, from the Department of Computer Science, at the University of Helsinki. In 1996, the Academy drafted the *Information Research Programme,* the purpose of which is to generate a multi-faceted approach to information, its production, presentation, transfer, and utilisation.

The aim of the consortium, *From Data to Knowledge* is to develop new methods for the analysis of large and complicated data sets, especially spatially referenced data. The consortium is based on the co-operation between computer scientists, mathematicians and statisticians, and other disciplines, including epidemiology, biotechnology, environmental sciences, geography, and archaeology. In the formation of knowledge from archaeological data, the aim is to combine GIS and spatial statistics. In particular, the Bayesian and Markov chain, Monte Carlo (MCMC), approaches are used. Investigation into uncertainty in data collecting, handling and analysis, and interpretation, is one of the main interests.

Uncertainty in archaeological data and interpretation

During the last years, the concept of *uncertainty*, differentiated from *error*, has gained a lot of attention in conferences and publications, especially in the fields of environmental research, decision making, and the use of GIS. Equally, the subject has been under discussion in archaeology (*e.g.*, Wiemer, 1995; Maschner, 1996; see also McGlade, 1997). Researchers have, of course, always been aware of the existence of uncertainty in their data and interpretation. However, the methods used for handling incomplete and imprecise data, as well as the epistemological basis, have changed with time. The increasing use of computer-based analysis has raised the problem of how to classify and handle uncertain information, in a binary language, without the words "maybe", "more", etc. In GIS studies, the difficulties in classifying continuous, or fuzzy, spatial data, without over-simplified classes or borders, has raised the need for non-classical analysis methods (see, *e.g.*, Burrough, 1996; Lagacherie, Andrieux & Bouzigues, 1996).

The problems are, how to cope with incomplete and imprecise data, how to describe this uncertainty, and how to measure its effects on results. When we talk about GIS, the aim is to use spatial data in advanced analysis, which demands much more from the data quality, than the pre-GIS tradition, based on paper maps.

In Fig. 1, archaeological data processing in GIS, has been described. Every GIS guide begins with the definition, that GIS is an information system, for collecting, storing, analysing, interpreting and presenting spatial data. In the same way, we can divide the causes of uncertainty and error, into these different levels of data processing. If we compare archaeological GIS to other disciplines, using GIS, what stands out for us is that the "real world", we are interested in, existed hundreds or thousands of years ago, and the "real world", we are observing, has undergone considerable changes, due to post-depositional formation processes. As a result, indirect observations cause uncertainty in interpretation and demands on theory: how do we interpret what we are observing?

Archaeologists are data producers, *i.e.*, they work on every level, from data collection to interpretation, by working with data, collected through archaeological excavations and surveys. However, archaeologists use a lot of environmental data (for example, soil maps), without adequate metadata, and information about the data quality.

P. A. Burrough (1987: 104) divided possible sources of uncertainty, in GIS, into three levels:

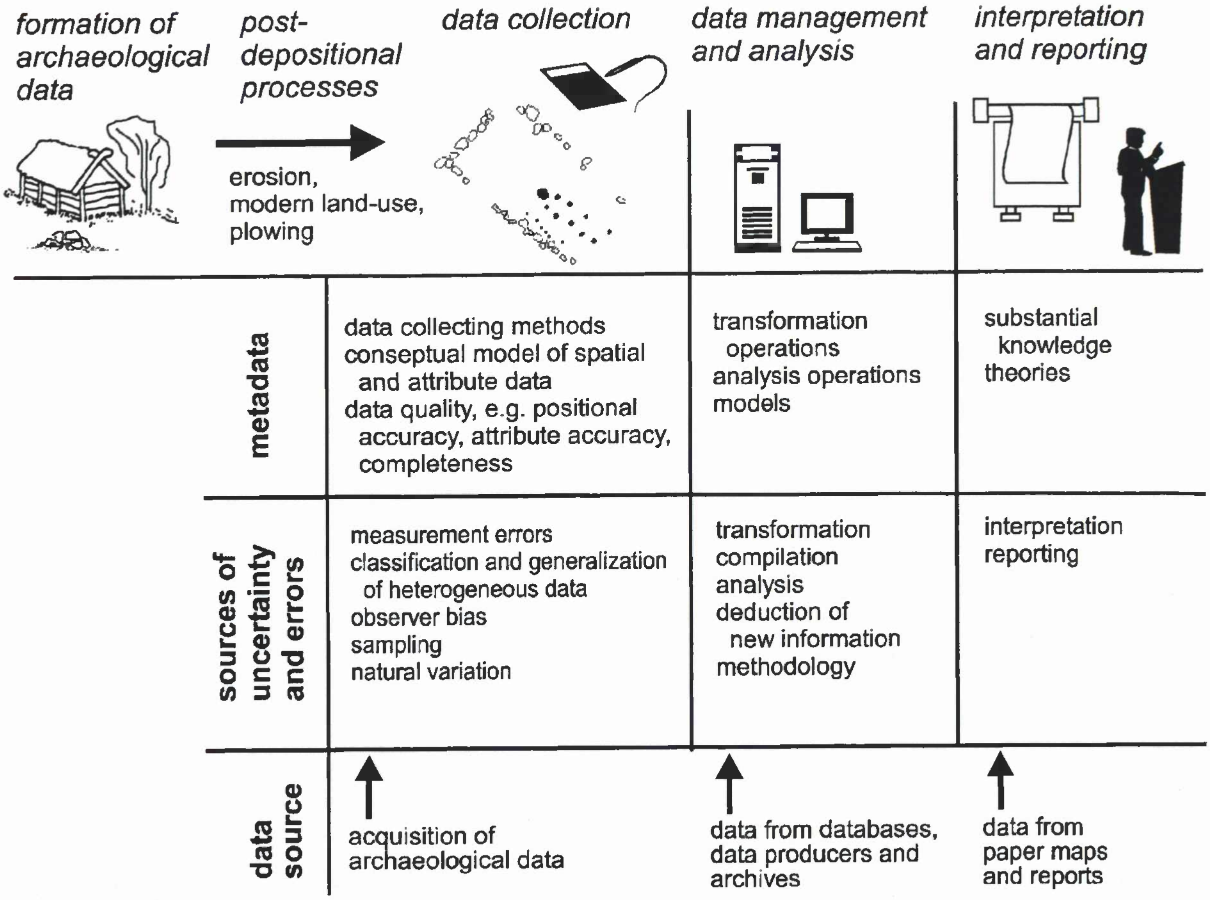
1. *Obvious sources of error*
   e.g., age of data, aereal coverage, density of observations, map scale

2. *Errors resulting from natural variations or from original measurements*
   e.g., quantitative and qualitative accuracy of content, sources of variation in data

3. *Errors arising through processing*
   e.g., classification and generalization problems: methodology, class interval definition, interpolation

According to Burrough, the first level sources of error are the most obvious, and they are easy to check (Burrough, 1987: 104). However, this kind of information is usually included in the metadata standards for digital information.

The second level contains more subtle sources of uncertainty and error, which can be detected, only while working intimately with the data. The accuracy of content depends much, on the decisions, classifications and generalizations, made by the researcher. For example, it is difficult to define the boundaries for continuous phenomena, e.g., soil types


formation of archaeological data
post-depositional processes
erosion, modern land-use, plowing
data collection
data management and analysis
interpretation and reporting

metadata	data collecting methods conseptual model of spatial and attribute data data quality, e.g. positional accuracy, attribute accuracy, completeness	transformation operations analysis operations models	substantial knowledge theories
sources of uncertainty and errors	measurement errors classification and generalization of heterogeneous data observer bias sampling natural variation	transformation compilation analysis deduction of new information methodology	interpretation reporting
data source	acquisition of archaeological data	data from databases, data producers and archives	data from paper maps and reports

Fig. 1. Uncertainty in archaeological GIS

(Burrough, 1987: 104), The important question is, how to make the information, used in decision making available as metadata.

At the third level, there is the uncertainty, arising through processing, for example, through interpolation. This level forms an important research area, for all scientists using GIS.

Examples of data processing

The research material consisted of two spatially referenced data sets. The first covered Iron Age sites and environmental factors, such as soil, water, digital elevation model (DEM), etc., over a large area, in Eastern Finland. The second was the material collected through the excavation and survey of an Eastern Finnish, Viking Age settlement site. The aim was to study the Iron Age settlement process, both at the regional level and at the local level.

The role of GIS, in this project, was to be a supplement to spatial statistics; in other words, it was used for storing, exploring, and visualizing spatial data. The procedure cuold be called, GIS-assisted modelling. The GIS-programs used were Idrisi for Windows and MapInfo.

Different methods, including fuzzy logic, were used to create map layers, that presented, not only observations, but also knowledge about the degree of uncertainty, in these observations. For example, while reconstructing the Viking Age water level in Eastern Finland, the estimation was based on the known, land-uplift rate, during the last 1000 years (see Nuñez, Vikkula & Kirkinen, 1995). Besides this, the yearly variation in water level, caused by climate, had to be taken into account. Because the variation could be predicted from present water stage measurements, it was possible to evaluate the degree of variation, in Viking Age water level estimation.

In Fig. 2, the map presents the estimated water level for the 9th century, AD. The average water level, normally used in map presentations, is marked in red-line and was measured, by taking the land-uplift rate into account. The effect of climatical variation is expressed, with different shades of green. The diagram below (Ympäristöhallinto 1997) shows the measurement information, from the last 80 years, which has been used for the evaluation of variation.

However, in many cases, the existence of uncertainty is acknowledged, but its exact effect on measurements and results is not known. For example, in Fig. 3, the aim was to determine the activity areas of a settlement site, by analysing soil profile observations (the presence of a coloured cultural layer or a stone structure) and phosphate contents.

The observations of a cultural layer were collected through sampling, and were classified using Boolean logic, while the phenomena was essentially continuous and fuzzy. Misclassifications were more common, when the cultural layer was light-coloured, or thin. Also, the variation, especially near the boundaries of the cultural layer, caused uncertainty in classification. The likelihood of errors, in classifications, was evaluated, by comparing the observations to reference material, collected through excavation.

Combining prior knowledge to spatially referenced archaeological data

The wide interest in combining relevant prior knowledge with the analysis, and the need to deal with error, have drawn attention to non-classical, statistical methods. The Bayesian method uses probability, as a means of measuring one's strength of belief, in the truth of a particular hypothesis. The Bayesian application of probability theory offers a well-founded method for handling uncertainty. The basic methodology, as well as a wide collection of archaeological case studies, have been presented in Buck, Cavanagh, and Litton (1996).

In our example of soil surveying, a model was formulated to express how soil data may reflect activity areas and land-use, at a typical Eastern Finnish Iron Age settlement site. The model consisted of rules describing the expected significance of soil profile observations and the results of phosphate analysis (see Table 1). The resulting probabilities, from a corresponding logistic regression model, are visualized in Fig. 4. The aim was to study the relationship between human activity intensity, expressed by soil samples, and, on the other hand, by excavations, while taking the prior knowledge into account.

In this project, the Bayesian inference was carried out by the Bassist system, developed at the University of Helsinki, Department of Computer Science, during the last three years. Bassist analyses computationally complex, Bayesian statistical models, using the so-called, Markov chain, Monte Carlo methods.

Conclusion

The exponentially increasing use of digital information has caused the need to study the basis and use of this information. In archaeological GIS studies, the combination of archaeological data bases and map bases require sophisticated methods, for handling uncertain and fuzzy data. Also, the importance of metadata, in data management, has become increasingly vital (see, *e.g.*, Farley & Gisiger, 1996; Wise & Miller, 1997).

This paper is a preliminary presentation of the ongoing project, *Spatial Data in Archaeology*, the purpose of which is to study and develop methods, for handling spatially referenced, archaeological data.

cultural layer observed in soil profile	cultural layer observed in neighbouring samples (#)	phosphate content	stone structure observed in soil profile	probability of archaeological features
0				0
1	0			0 – 0.1
1	1 – 2	< mean	0	0.1 – 0.3
1	1 – 2	> mean	0	0.2 – 0.4
1	1 – 2	< mean	1	0.2 – 0.4
1	1 – 2	> mean	1	0.3 – 0.6
1	> 2	< mean	0	0.3 – 0.5
1	> 2	> mean	0	0.4 – 0.6
1	> 2	< mean	1	0.5 – 0.7
1	> 2	> mean	1	0.7 – 0.9

Table 1. Decision rules for estimating human activity at an Eastern Finnish Iron Age settlement site through soil profiles and phosphate analysis. 0 = negative observation; 1 = positive observation.

References

BUCK, C.E., CAVANAGH, W.G. & LITTON, C.D. (1996), *Bayesian Approach to Interpreting Archaeological Data*, John Wiley & Sons, Chichester.

BURROUGH, P.A. (1987), *Principles of Geographical Information Systems for Land Resources Assessment,* Clarendon press, Oxford, 2nd.

BURROUGH, P.A. (1996), "Natural Objects with Indeterminate Boundaries", in: BURROUGH, P. & FRANK, A. (Eds.), *Geographic Objects with Indeterminate Boundaries. GISDATA* II, Taylor & Francis, London.

FARLEY, A. & GISIGER, A. (1996), "Managing the Infrastructure: The Use of Corporate Metadata for Archaeology", in: MASCHNER, H. (Ed.), *New Methods, Old Problems. Geographic Information Systems in Modern Archaeological Research,* Center for Archaeological Investigations, Southern Illinois University.

LAGACHERIE, P., ANDRIEUX, P. & BOUZIGUES, R. (1996), "Fuzziness and Uncertainty of Soil Boundaries: From Reality to Coding in GIS", in: BURROUGH, P. & FRANK, A. (Eds.), *Geographic Objects with Indeterminate Boundaries. GISDATA* II, Taylor & Francis, London.

MASCHNER, D.G. (1996), "Geographic Information Systems in Archaeology", in: MASCHNER, H. (Ed.), *New Methods, Old Problems. Geographic Information Systems in Modern Archaeological Research,* Center for Archaeological Investigations, Southern Illinois University.

MCGLADE, J. (1997), "GIS and Integrated Archaeological Knowledge Systems", in: JOHNSON, I. & NORTH, M. (Eds.), *Archaeological Applications of GIS: Proceedings of Colloquium II, UISPP XIIIth Congress, Forli, Italy, September 1996. Sydney University Archaeological Methods Series 5.,* Sydney.

NUÑEZ, M., VIKKULA, A. & KIRKINEN, T. (1995), "Perceiving time and space in an isostatically rising region", in: LOCK, G. & STANČIČ, Z. (Eds.), *Archaeology and Geographical Information Systems*, Taylor & Francis, London.

WIEMER, R. (1995), "Another way to deal with maps in archaeological GIS", in: LOCK, G. & STANČIČ, Z. (Eds.), *Archaeology and Geographical Information Systems*, Taylor & Francis, London.

WISE, A. & MILLER, P. (1997), "Why metadata matters in archaeology", *Internet archaeology*, 2, [http://intarch.ac.uk/journal/issue2/wise_toc.html]. Last updated Apr 30 1997.

Ympäristöhallinto (1997), Ympäristön tila. Mikkelin läänin vedenkorkeudet, December 22, 1997. [http://www.vyh.fi/tila/vesi/tilanne/wmik.html].

List of Figures in CD-ROM.

Fig. 2. Fuzzy water level at Lake Saimaa (Eastern Finland), in the 9th century AD. The diagram below: water stage measurement during the last 80 years, collected by Finnish Environment Institute (Ympäristöhallinto 1997). The yearly variation in water level is ±1.25 meters.

Fig. 3. Mikkeli Kenkäveronniemi, a Viking Age settlement site in Eastern Finland. Black dot = positive cultural layer observation; black circle = negative cultural layer observation; yellow star = positive observation of a stone structure; different shades of grey = phosphate content interpolation (white = 0-296 mg/kg,..., dark grey = 1000 mg/kg); red rectangle = excavation area. The distance between observations is 5 meters.

Fig. 4. Visualization of a logistic regression model of soil survey data (blue lines). Labels, see Fig. 3.