Archeologia e Calcolatori
7, 1996, 303-312

INFORMATION SCIENCE IN ARCHAEOLOGY: A SHORT HISTORY AND SOME RECENT TRENDS

1. Introduction

In the first section of this paper, I will try to sketch some of the developments in the use of information science in archaeology, putting them in a more general framework of developments in archaeological theory. I will discuss the shift from "classical" statistical approaches, which concentrate on hypothesis testing, towards more heuristic, pattern-searching methods of analysis. Besides showing my own biases, my emphasis will necessarily be on what happened in the "Anglo-Saxon countries" (Moscati 1990).

In the second section, I will discuss some research I am undertaking presently on the use of Bayesian statistics for solving archaeological problems. In doing so, I want to illustrate, on the one hand, the ease with which rather complicated quantitative analyses can be performed with the help of standard computing tools, and, on the other hand, the risks of carrying out such analyses without a clear, logically sound underpinning.

2. A short overview of the application of 'quantitative methods' in archaeology

2.1 *The beginnings*

When defining 'quantitative methods' as any type of recording or analyzing archaeological materials with the help of numbers, it can be stated safely that the use of such methods dates back to the beginning of the twentieth century. At that time large-scale, detailed excavations and detailed description of the finds and stratigraphy had come to be considered necessary for understanding the cultural history of the different groups of the past (Trigger 1989, chapter 5).

Whether these groups were seen as 'aliens', the case in North America, or as 'predecessors', as in Europe and the Near East, was in this respect unimportant. The detailed description of artifacts, needed for the construction of typochronologies, minimally involved counts, and the early attempts to seriate types by Petrie (1901) can be considered the first steps towards quantitative analysis. Most analysis, however, was restricted to the visual inspection and comparison of frequency tables, and not much changed in this situation for over fifty years.

2.2 *Hypothesis testing and numerical classification*

It is probably justified to say that the more rigorous application of in-

formation science to archaeology has most of its roots in the American 'New Archaeology', advocated by Lewis Binford in the early 1960s (e.g. BINFORD 1962, 1965). Under the banner of neo-positivism, the early New Archaeology set as its task the formulation of general laws and theories for human behaviour and the construction of models by which proposed laws and theories could be tested. To construct such models, detailed description, which puts the emphasis on variability, was neither sufficient nor adequate. Instead, the dimensions which represented the underlying general processes had to be made visible by removing all variability that was considered 'random' in relation to those processes (e.g. BINFORD, BINFORD 1966).

The emphasis on the utilization of archaeological data for testing general propositions led to extensive use of existing techniques of 'classical' inferential statistics, often without much attention being paid to the mathematical conditions that had to be fulfilled to validly apply such techniques. Moreover, in those days the available hardware and software often dictated which methods to use, instead of the archaeological problem setting.

Besides the American 'New Archaeology', another trigger for the development of applications of information science in archaeology was numerical classification. This approach, originally developed in biology (SOKAL, SNEATH 1963), appealed to many archaeologists on both sides of the Atlantic Ocean, and led to extensive studies and discussions of the vast gamut of methods that soon became available (e.g. HODSON 1969, 1970). It also triggered (again) the discussion of one of the basic problems in archaeology, the meaning and purpose of classification in general (e.g. DUNNELL 1971). As such, many applications of numerical classification by cluster analysis implicitly accepted the existence of 'natural' or 'objective' classes that could be unveiled by using an 'objective' method.

2.3 *Patterns and variability*

By 1980, the time of 'naive' applications of information science had ended. It had become clear that the variability in the archaeological record could not be directly explained by general laws and theories or, otherwise, safely could be neglected. The 'law and order' archaeology had grown into processual archaeology, partly rooted in the general theory of open systems, an approach which was followed by many, not only in the USA, but all over the world. Around this time, the interest in this type of archaeology also began to increase among classical archaeologists, who study Near-Eastern, Egyptian, Greek, and Roman archaeology. Without going back to a culture-historical approach, variability and the understanding of its causes again became a subject worth to be studied.

In the USA, the work of Michael Schiffer and his followers led to a much better perception of the many influences the archaeological record

undergoes before, during, and after its first formation. The archaeological record is *patterned*, and it is the task of the archaeologist to seek, describe, and explain the patterns. Explanations go from the level of understanding the local and general constraints the natural environment puts on the formation and preservation of the archaeological record to the level of understanding the impact, time-bound and spatially restricted, of human decision-making and action.

At this level, the level of 'middle range theory' (RAAB, GOODYEAR 1984), explanation seems to end, however. The next step--trying to find ahistorical and spatially unrestricted explanations, the goal of the early New Archaeology, was not made. This was perhaps not unexpected. Apart from the enormous amount of working and thinking required before this large-scale problem could, and can, be tackled, the unwillingness to address it correlated, in my opinion, with a general change in the outlook of Western society. Interest in investigating what people have in common became more and more replaced by interest in how people divide themselves under labels like 'cultural identity', 'nationality', and 'ethnicity'. In archaeology, this attitude, found its clearest expression in the so-called 'postmodern' approaches, which, fortunately, did and do not gain much influence.

In the area of classification, the search for the 'natural' order of things became replaced by the notion that any classification and ordering is only valid within the context of a well-defined research-design. The definition of classes, the selection of attributes, and the levels of measurement all are explicit decisions which reflect the purpose of the research undertaken (e.g. WHALLON, BROWN 1982; COWGILL 1990).

2.4 *Informed guesses*

I estimate that around 1985 the application of information science in archaeology entered a new phase, a phase we still find ourselves in. Describing and mapping the variability of the archaeological record in formal and quantitative terms has become familiar, and the construction of formal-mathematical models to develop theories that explain this variability now is one of the main concerns (e.g. DORAN 1990). Two approaches are of special interest.

The first concentrates on the construction of formal, dynamic models of processes of continuous change. These models are probabilistic in nature and tend to incorporate extensive computer simulations.

The second approach consists of attempts to understand behavioural processes in time and space by modelling human decision making with the help of concepts from artificial intelligence (e.g. DORAN, CORCORAN 1985; REYNOLDS 1986). In modelling decision making the Bayesian approach to statistical inference (HOWSON, URBACH 1993) is gaining more popularity (e.g. BUCK 1993). This important development recognizes that, even when design-

ing formal, 'objective' systems and processes, not all trajectories are initially equiprobable, or, even in case they are, will be recognized by the decision makers as such. Also, the Bayesian approach to information processing is, for me at least, intuitively satisfactory--it can serve as a formal model of the way in which the human brain human updates its knowledge and beliefs when new information becomes available.

2.5 *Between Archaeology and Information Science: The role of the 'quantitative archaeologist'*

To end the general section of this paper, I would like to discuss briefly the role of the archaeologist who has made the application of information science to archaeology his or her specialty. I did the same thing ten years ago, and, while many more archaeologists are now used to computers for storing and retrieving data and texts, I am afraid that on a more abstract level not too much has changed. It still is the case that only a minority of the archaeological community recognizes the potentials offered by information science for better understanding the archaeological record and, by that means, for better understanding humankind.

The metaphor I select to describe those of us who advocate a change in the thinking patterns of archaeologists is that of the *middleman.*«A 'classical' middleman is 'born' inside the culture of archaeology, has learned some of the language and culture of information science, alienates himself more or less from his archaeological culture and then functions as a channel through which information and goods are exchanged between both cultures. A middleman also has his own language. In that language, the concepts from different cultures are worded so that an interaction becomes possible. For applied information science in archaeology, the middleman language consists of mathematical models that are applied to archaeology.

Such models do more than only bridge the gap between archaeology and information science. They try to put concepts from archaeology and information science into a coherent, necessarily more general, framework of thought, thus creating concepts on a higher level of abstraction. Developing such concepts may be of more use to the 'parent' cultures than straightforward translations from one culture to another. In a different terminology: the historically determined dialectic opposition between science and humanities, as represented by old-fashioned mathematics and old-fashioned archaeology resolves itself in a synthesis--applied information science in archaeology.» (Voorrips 1985, chapter 1)

I think that what is expressed by the metaphor of the middleman still holds today. I also think that many of the prevalent attitudes and 'paradigms' in modern, or even later than modern, archaeology makes the work of the middleman extra hard, if not impossible. I hope, however, that it will be the

classical archaeologists I referred to above, the archaeologists who wrestle with the rich data set of the Mediterranean and the Near East, rich in all senses of the word, who will first understand the real meaning of information science for archaeology. They have a running start since they do not need to invent or reinvent many wheels--there already exists a vast amount of literature dealing with issues of methodology, including classification, sampling, modelling, simulation, systems theory, etc. Furthermore, the current state of computer technology, of both hardware and software, makes it possible to concentrate on the real issues without having to spend a lot of time circumventing technical problems.

3. An example of a Bayesian approach to (spatial) classification

In the second part of this paper I will describe some of my recent experiments with the application of a Bayesian approach to decision-making in the course of classifying spatial units into clusters on the basis of their attributes. The example is not typically archaeological, although the issue came up in the context of an archaeological problem setting.

In 1994, I spent a trimester at the Museum of Anthropology of the University of Michigan, Ann Arbor, USA, where I taught a graduate seminar on the application of Geographical Information Systems in archaeology. One of the data sets used was derived from the ongoing work of the director of the Museum, Professor John O'Shea, on the role agriculture played for the original inhabitants of the Northern part of the Michigan peninsula (O'Shea, Milner, in prep.).

One of the things my students and I tried to figure out was which types of natural forest-vegetation the inhabitants would have had to cope with in different locations. This is not an easy problem, because that part of the United States was completely deforested in the last half of the nineteenth century, and the vegetation types distinguished among the secondary growth in unexploited areas are supposedly rather different from the original ones.

There is, however, an interesting record with the help of which a reconstruction of the original forest communities could be attempted. This record consists of data collected around the mid of last century by the General Land Office. One of the tasks of this office was to perform a geodetic survey of the State of Michigan in order to construct cadastral maps. To that purpose, the geodesists, or 'chainmen' put markers at the corner points of every square mile and quarter square mile in the area. To be able to find back these markers later, the chainmen were instructed to record the species and the diameter of some nearby trees, as well as their direction and distance from the marker.

In general, four trees were described in this way at each corner point of every square mile, two trees at each corner point of every quarter square

mile, and, in addition, at least two trees along each mile-long section (BOURDO 1956). These trees were called 'bearing trees' or 'witness trees'. The chainmen were instructed to select the bearing trees on the basis of size and condition. This, together with a less-than-perfect knowledge of tree-species among the chainmen, and cases of obviously faked data, makes the sample to be found in the records of the General Land Office somewhat suspect. However, as various investigators of these records have reported, altogether the bias seems negligible (BOURDO 1956; HUSHEN *et al.* 1966).

The first step in using this sample of a little over 2500 trees to reconstruct the former forest-types in the region was to plot all tree locations on a map of the region. This map then was digitized, adding the information about the tree-type to each point. The digitized map and additional information were stored using the PC-GIS IDRISI (EASTMAN 1992).

Next, a decision had to be made on the definition of the spatial units to be used in the further analysis. One option was to put a grid over the area, and use the counts of the different species per grid-cell. This approach, which is technically the easiest one, has the drawback that it dissects the data, so that some trees located close to each other end up in different grid-cells. Another option, in my opinion the best one, would be to define a circle with a radius derived from the average distance of the bearing trees to the markers, to 'move' this circle over the map, and to establish non-overlapping spatial units whenever the number of points in the circle was over some threshold value. Unfortunately, no GIS-package or other computer program known to me at the time was able to do this (the paper by M. BAXTER and C.C. BEARDAH in this volume seems to provide a method to do this, however).

A third approach, and the one I decided to take, was to do a k-means cluster analysis of the locations, using their coordinates as variables. The number of clusters was set to 250, so that each cluster would contain approximately ten points. The coordinate-data were transferred from IDRISI to SPSS/PC+, and put through the k-means procedure this package provides (NORUSIS 1988). The k-means procedure of SPSS is rather primitive compared to the one found in Kintigh's package 'The Archaeologist's Analytical Toolkit' (KINTIGH 1988). One manually has to repeat the analysis until a stable solution has been reached, and there is no provision to compare the solution with solutions based on randomized distributions of the data, as available in Kintigh's package. However, the distance of each point to the centre of the cluster it belongs to, data needed later in the analysis, is included in the standard output.

The coordinates of the centres of the 250 clusters found were transferred back to IDRISI and translated into a point map. Using one of the IDRISI-routines, I constructed Thiessen polygons around the cluster centres and used these polygons as the spatial units in the rest of the analysis. Next, I wrote a small Fortran program that aggregated the information on the tree-type of each point and the spatial cluster it belonged to into a table of 250

rows, the spatial clusters, and 33 columns, the number of different tree-types. The cells of this table contained the counts for the tree-types, and, as can be imagined, most of them were empty. The table I transferred to a database management package for PC, in this case Microsoft Access.

At this point in the analysis it was necessary to decide on the method to use to group the spatial units into something that might represent the former forest-types. An obvious candidate for clustering the spatial units was again the k-means clustering procedure using the different tree-types as variables. But first, the problem of the empty cells had to be tackled. Without solving this problem, the choice would be either obtaining clusters composed of locations that shared the *absence* of many tree-types, or using a similarity coefficient with questionable mathematical properties, such as Jaccard's coefficient. After looking at the frequency distribution, I first removed all tree-types that occurred less then 40 times, ending up with 2333 points divided over 12 types, which was, on average, still less then one per cell. I then used Kintigh's k-means procedure, computing the solutions for two to ten clusters, and checking the validity with the randomization test.

The results were not convincing since there was only a weak patterning in the data. As for the number of clusters to be distinguished, five seemed to be the best. I added the results of the five-cluster solution to the table in the database.

It was clear that the matrix was too sparse for cluster analysis of some kind or another, and I therefore decided to try a different approach, inspired by the ideas of, among others, C.D. Litton (Buck, Litton 1993). I first assumed that there indeed were five forest-types in the region. Given that forest-types had been distinguished elsewhere by botanists trying to reconstruct the old forest vegetation in other parts of Michigan, I then assumed that it would be possible to decide on botanical and pedological grounds which five of those types could be expected to have existed in the study-area, may be assisted by the results of the k-means cluster analysis, weak as they were. If the probabilities of the occurrence of the different tree-types in those five forest-types were known, then, by applying Bayes theorem I could estimate the probability that a spatial unit belonged to one of the vegetation types, looking at the tree-types the unit contained. In that manner I would use only the 'real' information in the data and not similarities based on the empty cells in the matrix.

Unfortunately, it soon became clear that more or less reliable estimates for the probability of a tree-type to occur in a vegetation-type were not in the literature. The only possibility for estimating these probabilities was to make one more assumption, which was that the results of the k-means cluster analysis could be taken as a rough approximation of the composition of the former forest-types in the region. Under that assumption, I could compute the probabilities with which each tree-type occurred in each of the five forest-types.

Using the database I aggregated the data necessary, after which I had to

write a small Fortran program to perform the actual computations.

I decided to use only those points in each unit which were at less than average distance from the cluster centre, assuming that in most cases these would be enough to reach an unambiguous result. By doing so, the compactness in space of the points used would be high, hopefully leading to better defined vegetation units. To calculate average distances and select the points obeying the criterium, I created a database table containing the identifiers of the spatial units and the distances of each point toward the centre of the unit it belonged to, using the output from the SPSS locational clustering. A simple query of this table produced the selection.

The outcomes of the analysis were interesting. A large number of the spatial units had a final probability of over 0.9 of belonging to a specific forest-type, and in a majority of the cases a specific forest-type had a probability of over 0.7. I decided to use 0.7 as cut-off value so that units with a lower probability for any of the forest-types were considered unclassified. To date, this is as far as the analysis has gone.

What can be done with the unclassified units? Here, again, a Bayesian approach may be useful. New sets of posterior probabilities can be calculated for them using, one at a time, the final probabilities of their neighbours (classified or not) as new information. This will, depending on the neighbour selected, in most cases result in one of the five posterior probabilities reaching a value of over 0.7. Identification then will be into the forest-type with the highest posterior probability. If no posterior probability reaches a value of over 0.7 the unit can be considered to represent a transition zone between two forest types.

The main reason for this rather detailed description of a small part of a project that is still under construction and, probably, of very restricted scientific meaning is to show how a Bayesian approach can offer alternatives for solving practical problems. A second reason is to show that the process is far from automatic--almost at every stage decisions by the investigator on how to proceed are necessary.

Finally, I wanted to show the relative ease with which an investigation like this one can be done using a few computer tools. Without a GIS, the calculation of Thiessen polygons is a nightmare, without a package like SPSS/PC+ or Kintigh's Toolkit, k-means cluster analysis is impossible, and without a relational database management package many manipulations of the data are cumbersome, to say the least. Notwithstanding all the ready-made computer tools, however, it was still necessary to write a few simple computer programs myself, to perform a number of needed computations not included in the available packages.

Albertus Voorrips
Faculty of Environmental Sciences
University of Amsterdam

Acknowledgements

First of all, I want to thank the organizers, in particular Dr. Paola Moscati, for inviting me to participate in the *III Convegno Internazionale di Archeologia e Informatica*, in Rome, November 1995. I am very grateful to the Museum of Anthropology, University of Michigan, which made it possible for me to spend a fruitful trimester in Ann Arbor in the fall of 1994, during which the research described in the second part of this paper was initiated. Special thanks go to the students who attended the seminar on 'GIS in Archaeology' I taught at that time. Their enthusiasm and eagerness provided me not only with a lot of tediously computerized raw data, but with much excellent food for thought as well. A final word of thanks goes to Susan Holstrom who corrected the English.

BIBLIOGRAPHY

BINFORD L.R. 1962, *Archaeology as Anthropology*, «American Antiquity», 28, 217-225.

BINFORD L.R. 1965, *Archaeological Systematics and the Study of Culture Process*, «American Antiquity», 31, 203-210.

BINFORD L.R., BINFORD S.R. 1966, *A preliminary analysis of Functional Variability in the Mousterian of Levallois France*, in J.D. CLARK, F.C. HOWELL (eds.), *Recent Studies in Paleoanthropology*, «American Anthropologist», 68, 238-295.

BOURDO E.A. 1956, *A Review of the General Land Office Survey and of its Use in Quantitative Studies of Former Forests*, «Ecology», 37(4).

BUCK C.E. 1993, *The provenance of archaeological ceramics: a Bayesian approach*, in J. ANDRESEN, T. MADSEN, I. SCOLLAR (eds.), *Computing the Past: Computer Applications and Quantitative Methods in Archaeology, CAA92*, Aarhus, Aarhus University Press, 293-301.

BUCK C.E., LITTON C.D. 1993, *Application of the Bayesian paradigm to archaeological data analysis*, in J. PAVÚK (ed.), *Actes du XII^e Congrès International des Sciences Préhistoriques et Protohistoriques 1*, Nitra, Institut archéologique de l'Académie Slovaques des Sciences à Nitra, 367-374.

COWGILL G.L. 1990, *Artifact classification and archaeological purposes*, in A. VOORRIPS (ed.), *Mathematics and Information Science in Archaeology: a Flexible Framework*, Studies in Modern Archaeology 3, Bonn, Holos Verlag, 61-78.

DORAN J., CORCORAN G. 1985, *A computational Model of Production Exchange and Trade*, in A. VOORRIPS, S.H. LOVING (eds.), *To Pattern the Past*, PACT 11, Strasbourg, Council of Europe, 349-359.

DORAN J.E. 1990, *Computer-based simulation and formal modelling in Archaeology: a review*, in A. VOORRIPS (ed.), *Mathematics and Information Science in Archaeology: a Flexible Framework*, Studies in Modern Archaeology 3, Bonn, Holos Verlag, 93-114.

DUNNELL R.C. 1971, *Systematics in Prehistory*, New York, The Free Press.

EASTMAN J.R. 1992, *IDRISI, User Guide and Technical Reference Manual*, Worcester, Clark University.

HODSON F.R. 1969, *Searching for Structure within Multivariate Archaeological Data*, «World Archaeology», 1, 90-105.

HODSON, F.R. 1970, *Cluster Analysis and Archaeology: some new developments and applications*, «World Archaeology», 1, 299-320.

HOWSON C., URBACH P. 1993, *Scientific Reasoning: the Bayesian Approach*, second edition, Chicago and La Salle, Illinois, Open Court.

HUSHEN T.W., KAPP R.O, BOGUE R.D., WORTHINGTON J. 1966, *Presettlement Forest Patterns in Montcalm County, Michigan*, «Michigan Botanist», 5, 192-211.

KINTIGH K. 1988, *The Archaeologist's Analytical Toolkit (C) 1985-1988*, Tempe, Kintigh.

MOSCATI P. 1990, *Indirizzi e sviluppi dell'Archeologia Quantitativa*, in P. MOSCATI (ed.), *Trattamento di Dati negli Studi Archeologici e Storici*, Roma, Bulzoni, 1-54.

NORUSIS M.J., SPSS Inc, *SPSS/PC+, Advanced Statistics V2.0*, Chicago, SPSS Inc.

O'SHEA J.M., MILNER C.M. in prep., *Late Prehistoric Economy and Ecology in Northeastern Lower Michigan*, «Memoirs of the Museum of Anthropology», Ann Arbor, University of Michigan.

PETRIE W.M.F. 1901, *Diospolis Parva*, London, Egypt Exploration Fund.

RAAB L.M., GOODYEAR A.C. 1984, *Middle range theory in archaeology: a critical review of origins and applications*, «American Antiquity», 49 (2), 255-268.

REYNOLDS R.G. 1986, *An adaptive computer model for the evolution of plant collecting and early agriculture in the eastern Valley of Oaxaca*, in K.V. FLANNERY (ed.) *Guila Naquitz: Archaic Foraging and Early Agriculture in Oaxaca, Mexico*, Orlando, Academic Press, 439-500.

SOKAL R., SNEATH P. 1963, *Principles of Numerical Taxonomy*, San Francisco, Freeman.

TRIGGER B.G. 1989, *A History of Archaeological Thought*, Cambridge, Cambridge University Press.

VOORRIPS A. 1985, *Chancy Choices: mathematical models in archaeology*, PhD thesis, Amsterdam, University of Amsterdam.

WHALLON R., BROWN J.A. (eds.) 1982, *Essays on Archaeological Typology*, Evanston, Illinois, Centre for American Archaeology Press.

ABSTRACT

In the first section of this paper, some of the developments in the use of information science in archaeology are discussed, putting them in a more general framework of developments in archaeological theory. It shows the shift from "classical" statistical approaches, which concentrate on hypothesis testing, towards more heuristic, pattern-searching methods of analysis. In the second section, some research is presented on the use of Bayesian statistics for solving archaeological problems. It illustrates, on the one hand, the ease with which rather complicated quantitative analyses can be performed with the help of standard computing tools, and, on the other hand, the risks of carrying out such analyses without a clear, logically sound underpinning.