36 Pre–processing of archaeological data

Jan Rulf

36.1 DATA PRE–PROCESSING

According to M.K. Chytil (1983, 1984), data pre–processing constitutes the first of three parts of data analysis — data pre–processing, algorithmical analysis, and interpretation. The aim is to furnish the archaeologist with appropriate data (i.e. the data are required to be semanticly representative, correct and contextual). The pre–processing begins with formulating the problem and results in the data appropriate for subsequent analyses. Data pre–processing consists of preparatory (planning and data acquisition) and filtering procedures.

36.2 ARCHAEOLOGICAL DATA

Archaeological research is often thought to be merely a question of collecting data, where the central entities "find" and "date" are identified as part of the process. This, however, is not the case: archaeological data arise as a result of a specific evaluation of information obtained in the field or from other sources. Data should be understood as a value of variable quantity (attribute), representing the state of some quality (variable) of the object investigated (thing, individual, system; the term object includes everything that can be speculated upon). A decision as to what variable values shall be classified for the object investigated is the so–called "standard". There are also other definitions of data. For J.W.Tukey (1980), e.g., data are codes representing the characteristics, measuring values or replies to questions.

According to E. Neustupný (1986, 1989), basic terms in archaeological theory are the entity (a find, e.g. a vessel) and the quality (the entity's quality, e.g. the form of the vessel). In my opinion, entities can be either concrete/particular (real existing things, persons, relations, systems), or ideal (concepts of concrete things or their combinations), or formal, lacking real counterparts (i.e. e.g. formal artefact types). In the above mentioned definition, qualities are represented by means of variables, the data reflect the states of these variables (e.g. the vessel form: B5). In archaeological theory, the qualities (sometimes called attributes — *Merkmale*) can be divided into external (also called control/evidence qualities (Pavlů1981), manifesting relations between entities, e.g. the situation of a vessel within the investigated area), and internal ones (i.e. diagnostic, expressing the actual quality of an entity, e.g. colour). Within the division into discrete (nominal) and continuous variables (ordinal, interval, ratio), archaeology mostly deals with nominal variables. There is of course the possibility of transforming the variables, e.g. when nominal variables are defined through calculations of ratio scale variables (Shennan 1988). Also, the arrangement within the frame of nominal variables can be understood as the classification itself (Spaulding 1982). During the analysis, the variable is sometimes separated from the relevant quality, resulting in a new ideal variable (especially in multivariate analysis). In this case the feedback decoding of variables to specific qualities or their complexes is a matter of interpretation.

36.3 PRE–PROCESSING OF ARCHAEOLOGICAL DATA

36.3.1 Planning of data

36.3.1.1 Problem formulation

This is the basic step that states the aim to be reached by means of data. This can be, e.g., museum evidence of a find, description of an object for a popular explanation; in most cases, however, it is the solution of a specific scientific problem. The formulation starts with realisation of lack of knowledge in a specific context. A further step is the posing of problem, i.e. the explicit expression of lacking knowledge and the formulation of the problem, i.e. expression of the question that shall be answered. At the same time, background knowledge on which the solution is based must be delimited.

36.3.1.2 Scheme of the data model

This scheme starts with the formulation of the problem. Its creation means actually the selection of entities (of a counted quantity yielded by excavation or in another way), selection of qualities, i.e. the descriptive space of qualities, the number of which can be theoretically infinite, and the selection of eligible variables, which should represent the states of qualities. This selection, depending merely on our decision, brings a subjective element into the data model, which is under the sway of a specific paradigm (this is also the viewpoint held by C. Orton 1980: «The data are a record of the relationship between the recorder and the recorded»). The data model has to answer a difficult question, i.e. how many and what data do we need for the solution of the problem stated. This task is in close association with the complicated and important problem of sampling in archaeology, where one of the basic questions is what constitutes a sufficient size of a sample (Shennan 1988). Last but not least, while building the data model we should also consider the demand of minimum time and costs needed for the solution of the problem. Thus, the data model consists of one or more descriptive systems. The systems are either primary (their application is based on mere observation or measurement) or secondary (according to E. Neustupný 1986; the latter arise as a result of multiplication of two primary systems mutually transposed — e.g. the matrix of correlation coefficients). Creation of data models for the solution of specific problems should be iterative, since our ideas concerning the importance of individual entities and qualities are defined more precisely only after viewing the results of the analysis. Besides, creation of descriptive systems is rather time–consuming (Pavlů 1981). Also exploratory analyses of specific data samples are very useful. They help to clear up the behaviour of individual qualities within various structures. Another important aspect of data model building is the estimation of qualities, and the determination of their hierarchy.

The creation of data models, as well as the next phase — data acquisition — brings about many potential mistakes, resulting from imprecise, unstable and ambiguously defined terms. The requirement for the archaeological description language — to be as formal as possible (and to differ from the interpretation language — documentation language or meta–language according to Gardin 1979) — is thus not fulfilled. The definitions of specific qualities are expected to be objective, unbiased, logical, independent, standard, exact, and to have adequate refinement (Malina 1977). Numerous attempts to reduce the lack of restraint in archaeological language, and to unify the terminology to a certain degree, at least on a common regional–chronological basis, have not brought any conspicuous results so far.

Another possible source of mistakes, arising already during the data model building, is the lack of unified methodology in entity and quality description. This is caused by an insufficient level of development in archaeological theory, resulting in the incompatibility of individual data models to the solution of similar scientific problems, and in the impossibility of mutual control of interpretation results. In fact, it is not as difficult to establish specific description norms, as it is to reach general agreement and acceptance of this norm (here we meet with similar difficulties as with the unification of terminology).

36.3.2 Data acquisition

Data acquisition can be divided into two basic procedures: empirical observation and formal evaluation. Here too other scientific disciplines should assist archaeology, e.g. mathematics, logic, measurement theory, linguistics. In general, the data should be exact enough to match their purpose (within the theory of measurement, exactness is a special relation between two different sets, expressing the planned stage of concord of two entities: data exactness should also be considered in the data model building). Further, data should be correct (correctness is a pragmatic term concerning the concord between the presupposed and actually achieved result). Most frequent sources of data in archaeology are observation and measurement, less frequent are experiments and monitoring.

The proper acquisition of archaeological data represents a source of many mistakes itself, since objective evaluation of the states of investigated qualities is often missing. Materialisation of the description of nominal variables, a guarantee for disjunction and identification independent of the observer, is extremely difficult, and often the qualities have to be further decomposed to sub–qualities represented by measurable variables (e.g. for the evaluation of pottery it is the degree of firing, size and kind of tempering, mineralogical analysis of the pottery material, etc.).

Data acquisition is the application of descriptive systems in practice. During this process, mistakes have the following causes (cf. Podborský *et al.* 1977):

1) the description is carried out by different persons, whose knowledge of material and approach to the description need not be the same;
2) the borderline between nominal variables can be wrongly determined by individuals;
3) states of variables are measured using different instruments and devices working with specific errors;
4) incidental mistakes (omission, clerical errors, typing errors).

36.3.3 Filtering procedures

Individual steps of this procedure can be described as follows:

- checking for and deleting defect data;
- organisation of data models;
- description of data, transformation of data; and
- information about the data.

In general, data defects can be described as follows (cf. Chytil 1983):

1) in most cases, data defects are contextual or problem–dependent, i.e. they are not absolute;
2) each data complex is defective to a certain degree;
3) results of any statistical procedure are defective if based on defective data;
4) it is better to deal with the defects right at the beginning than to eliminate them during the analysis;

In addition to the discovery, deletion and replacement of errounious data, this procedural stage includes also the establishment of new, transformed data (e.g. the data of secondary descriptive systems, or transformation of data without common classification, etc.), and the information about the data themselves. Here, the visual or numerical summaries of a single variable (Shennan 1988) will be considered first of all. In visual summaries, it is especially the histograms, column and pie graphs, cumulative curves, etc. that are of interest. In numerical summaries it is the descriptive statistics: mean, median, skewness, kurtosis, dispesion etc. These data characteristics provide a basis for the further selection of computer strategy, since for example, the application of various statistical tests depends on the data lay–out.

36.3.4 Data pre–processing and statistical software

As previous research has demonstrated (e.g. Chytil 1984), only a part of the problems associated with data pre–processing can be algorithmised and solved by computers.

In the future, specialised expert systems might be of a great importance in problem formulation and data model building. Recording all the already known standard solutions of selected problems and all well–tried data models, they may improve the orientation in background knowledge. The database systems are of course most suitable for proper data manipulation. In addition to common systems, like e.g. dBase IV, the CLIO system specialised in social sciences should be mentioned (cf. Trenkler 1990), or the older American system ADAM (Archaeological Data Management — Gaines 1981). Also large statistical packages like BMDP, SAS and SPSS enable modification and filtering of data complexes by appropriate variable formatting, further computation of new variables (command "compute"), recording of new variable values ("recode"), definitions of missing values ("missing values"), limitation of data selection in certain cases ("select if"), data weighting ("weight by"). One of the few specialised archaeological software means, The Bonn Seriation and Archaeological Statistic Package, does not include any special procedures to facilitate data pre–processing.

36.4 CONCLUSION

Pre–processing of data represents an integral part of the archaeological analysis. Demands for the quality of the data grow together with the number of finds and the increasing hardware possibilities of computer technology, while no appropriate attention has been devoted to pre–processing and the whole philosophy of the creation of data complexes. However, it is a major task for archaeological theory to solve the questions associated with the problem formulation and the data model building. In my view, it is just the imperfection of archaeological data pre–processing that hinders effective application of computers and quantitative methods in archaeology.

References

Chytil, M.K.
1983 Data Pre–processing and Computational Analysis, *Statistical Software Newsletter*, Vol.9,No.1:3–16.

1984 Is data pre–processing a computational process only?, *COMPSTAT* 1984:467–472.

Gaines, S.W. (ed.)

1981 *Data bank application in archaeology*. Tuscon.

Gardin, J.–C.

1979 *Une Archéologie Théoretique*. Paris.

Malina, J.

1977 *System of Analytical Archaeography*. Praha.

Neustupný, E.

1986 Nástin archeologické metody — An outline of the archaeological method, *Archeologické rozhledy* XXXVIII:525–549.

1989 Comments on archaeological data, in: *Bylany–Seminar* 1987. Collected papers, Praha, pp. 45–47.

Orton, C.

1980 *Mathematics in Archaeology*. Cambridge.

Pavlů, I.

1981 Die Deskription der Linearbandkeramik: Möglichkeiten und Grenzen, AFD Beiheft 16, *Beiträge zur Ur–und Frühgeschichte I*:145–150.

Podborský, V., Kazdová, E., Koštuřík, P. & Weber, Z

1977 *Numerický kod moravské malované keramiky. Numerische Kode der mährischer bemalten Keramik*. Brno.

Shennan, S.

1988 *Quantifying Archaeology*. Edinburg.

Spaulding, A.C.

1982 Structure in archaeological data: nominal variables. in R. Whallon & J.A. Brown (eds.) *Essays on Archaeological Typology*. Evanston pp. 1–20.

Trenkler, C.

1990 Altertumswissenschaft und historische Fachinformatik, *Archäologische Informationen* 13, H.2:121–126.

Tukey, J.W.

1980 Styles of Data Analysis and their Implication for Statistical Computing, *COMPSTAT* 1980:21–31.

Author's Address
Archeologický ústav ČSAV
Malá Strana, Letenská 4
CS–118 01 Praha 1