Fifteen years of contributions of the French school of data analysis to quantitative archaeology

F. Djindjian

20.1 Introduction

In France, at the beginning of the 1970's, several statistics laboratories contributed to a quantitative movement today known as 'the French School of Data Analysis'. The most widely known results are numerous multivariate analyses, including the famous Correspondence analysis, and many cluster analysis algorithms.

The aim of this paper is to present a synthesis of the archaeological applications of data analysis in France and to place it in a more general context of data analysis in Archaeology.

French contributions to the progress of quantitative methods in Archaeology, such as:

- Typological Analysis,
- Morphological Analysis,
- Culture pattern studies,
- Seriation and Toposeriation,
- Intrasite Spatial Analysis,
- Provenance studies,
- Palaeoclimatology

and their possible integration in a more or less explicit cognitive process are given as well.

20.2 The French school of data analysis

After 1965, in the universities and in the research institutes, in France, statistical programs were developed, dealing with multivariate data analysis. Namely:

- J.P. Benzecri and his team of the laboratory for Mathematical Statistics, in the University of Paris VI (with B. Escofier, J.P. Fenelon, P. Cazes, M. Jambu, M. O. Lebeaux, L. Lebart, M. Roux ....) They conceived the term 'Analyse des Données', and developed numerous algorithms including Correspondence Analysis;
- I.C. Lerman and his group in the University of Rennes, working in the field of the cluster analysis;
- E. Diday, in the INRIA-CNRS (Research Institute in Cybernetics and Computers), working in the field of non-hierarchical cluster analysis.

These major teams paralleled the research that was being done in applied research laboratories, namely:

* CNRS U.A. 880 (Musee des antiquites nationales)
and between 1965 and 1975, in the field of Archaeology, F. de la Véga, A. Guenoche and Ph. Cibois in M. Borillo’s LISH Laboratory (Laboratoire d’Informatique pour les Sciences Humaines).

The vitality of the French School of Data Analysis has been characterised by an intensive research activity in the seventies, resulting in:

- Data Analysis fundamentals (Benzecri, 1973; Lerman, 1970),
- Correspondence analysis (Escofier, 1969), Principal Component Analysis, ranking order factor analysis, time series factor analysis, PCA on categorical data, multiple factor analysis, discriminant analysis on categorical data, non linear multiple regression, etc...
- Hierarchical ascending cluster analysis, with fast algorithms, cluster interpretation ratios, correlations with factor analysis, aggregation criteria based on cluster variance, tests and simulations for dendrogram interpretation, non-hierarchical cluster analysis, cluster analysis on variables, etc... (Jambu and Lebeaux, 1983), (Lerman, 1981), (Diday, 1980).

The spread of the data analysis algorithms has been associated with the distribution of data analysis packages, the most known of them being:

- LADDAD (Laboratoire de Statistique Mathématique)
- SPAD (CREDOC).

20.3 The applications of data analysis techniques in archaeology

20.3.1 The introduction of Data Analysis in Archaeology took place simultaneously in 1966 in the U.K following the impact of the book of Sokal and Sneath (1963): Numerical Taxonomy, by two papers: Hodson, Sneath and Doran (1966), Doran and Hodson (1966), and in the U.S, following the use of factor analysis in psychometry: Binford and Binford (1966).
20.3.2
The spread of data analysis in Archaeology was very fast between 1966 and 1970, culminating in two large proceedings of symposia:

- 'Archéologie et Calculateurs' in Paris (Gardin, 1970),
- 'Mathematics in the Archaeological and Historical Sciences' in Mamaia (Hodson, Kendall, Tautu, 1971).

The principal techniques of interest were Multidimensional Scaling and Cluster Analysis, as well as similarity and distance measures. Principal Component Analysis replaced factor analysis, which really has never been applied to Archaeology, and then progressively multidimensional scaling techniques.

Numerous applications of Data Analysis techniques took place in the U.K inspired by the work of Doran and Hodson; in France, following the research of the L.I.S.H. with Borillo, de la Vega and Guenoche, and in many universities across the U.S.

20.3.3
The first half of the seventies witnessed a broader use of cluster analysis in other European countries: Belgium (Cahen and Martin, 1972), Italy (Bietti, 1978), Germany (Bergmann, 1970; Hahn, 1976).

20.3.4
In 1975, a first synthesis of Data Analysis in Archaeology was published: Doran and Hodson (1975), 'Mathematics and Computers in Archaeology'. This text emphasized the technical revolution of Data Analysis in the field of Typology, Culture pattern studies and Seriation.

20.3.5
The Correspondence Analysis was first applied to Archaeology in France from 1975, following the publication of the book of J.P. Benzecri (1973) in a general movement of enthusiasm for the application of data analysis techniques. The first applications, between 1975 and 1980, were:

- Culture pattern studies and seriation: Djindjian (1976a, 1977b), Hours (1976), Laplace and Merino (1977), Bergougnan and Mohen (1980);

A first synthesis of different applications to Archaeology was published in 1980 (Djindjian and Leredde).

20.3.6
The eighties saw a broader spread of multivariate analysis (Principal Component Analysis and Correspondence Analysis) in the European countries that previously employed only cluster analysis, namely:

- Germany: Stehli and Zimmermann (1980),
- Spain: De Quiros (1982),

In Eastern Europe, in the Soviet Union (Dolukhanov, Kozlowski and Kozlowski, 1980) and in Poland (Schild, 1979; Burdukiewicz, 1981), archaeologists discovered the rise of Cluster Analysis and Principal Component Analysis.

20.3.7

The introduction of Correspondence analysis to the Anglo-Saxon countries has been delayed in spite of the paper of Hill (1974) giving the main features of the technique, and the close link with the reciprocal averaging method.


20.4 The influence of data analysis on the evolution of archaeological methods

Data Analysis techniques have made a major contribution to almost all archaeological methods from the seventies. Giving to archaeologists the capability to avoid the fastidious computations of elementary or classical statistics, these techniques allowed them progressively to build a more formalised methodology.

20.4.1 Typology

Empirical typologies, matrix analysis, attribute association tests (Spaulding, 1953), all led to the first cluster analysis applications (Hodson, Sneath, Doran, 1966). From 1966 to 1975, Doran and Hodson (1975) tested different Q-mode data analysis techniques: Multidimensional scaling, Principal Component Analysis, K-means analysis, Discriminant analysis, Constellation Analysis. The fundamental role of the variables in the variance pattern was emphasized by Djindjian (1976a, 1977a) who introduced the simultaneous use of the R + Q mode Correspondence analysis and a variance based ascending hierarchical analysis, on tables of raw data transformed by a disjunctive coding (0,1). The need for an exhaustive description of artefacts was abandoned for the research on a minimum homogeneous intrinsic description (Djindjian, 1980b), improving the evidence and the stability of the clusters, and allowing extrinsic and intrinsic cluster interpretations. Separate processing of different intrinsic descriptors, with their respective coding, was proposed by Decormeille and Hinout (1982) in a new concept of multiple typological analysis.

20.4.2 Morphology

The understanding of the specific issues in shape analysis, (for example, in ceramic studies), and consequently with classical taxonomic approaches was due to U.K archaeologists from 1975: Wilcock and Shennan (1975) tested the efficiency of the mosaic method and of the sliced method while, in France, the ratio based methods were preferably used: (Djindjian and De Croisset, 1976), Guenoche and Tchernia (1978), Mohen (1980).

Recent studies (for example in Djindjian ed, 1985) have shown the main influence of the following three factors:

- the variance, and then the discriminant efficiency of a coding, for given type of forms,
• the capability of data analysis techniques (PCA, CA), applied to a given coding, to separate the size factors from the shape factors,

• the necessity to adapt iteratively a coding, in place of searching an optimal coding.

At the same time, the use of automated digital measure tools (Kampfmeyer, 1986) does not conceal the difficulties due to the discriminant performances of a form coding, which can be only ameliorated by iterations.

20.4.3 Intrasite Spatial Analysis

Single spatial distribution analysis (for example, NNA test), association tests between pairs of spatial distributions (for example, permutation test), all led to the rise of data analysis techniques, based on the concept of a local multivariate density measure.

Johnson (1977) and Graham (1980) proposed a R-mode multivariate analysis, applied to a N.N.A-like similarity matrix. Kintigh and Ammermann (1982) used a cluster analysis directly on the (x,y) coordinate matrix of each spatial distribution. However this method of looking for the superpositions of clusters cannot provide a real multivariate intrasite spatial analysis. Hesse (1984) and Bouchet (1986) used correspondence analysis directly on the counting cell matrix.

Whallon (1985) developed a technique termed: 'Unconstrained Clustering' based on a cluster analysis applied to a multivariate local density matrix, after having smoothed the spatial distributions.

Djindjian (1988), proposed an improvement of Whallon's method with the simultaneous use of a correspondence analysis (as a R+Q mode non linear multivariate technique) and a hierarchical cluster analysis with spatial proximity constraints.

20.4.4 Culture pattern studies

Apart from the R-mode factor analysis of Binford (1966), the applications of cluster analysis, multidimensional scaling (Doran and Hodson, 1966), PCA (Hodson, 1969), Canonical analysis (Graham, 1970), and Constellation analysis (Azoury, Hodson, 1973) are all Q-mode analysis.

Correspondence analysis, associated with hierarchical cluster analysis, has been applied as a R+Q mode analysis with a X2 metric well adapted to abundance tables allowing the determination of the role of different types in the patterns.


The method has been extended to a R+Q mode analysis on attribute tables (called Burt tables) by Djindjian (1980a) and integrated into a general methodology for intrinsic and extrinsic interpretations of culture patterns (Djindjian, 1980b, 1985).

20.4.5 Seriation

Ordering techniques by a direct (0,1 tables) or an indirect (similarity table) processing, have been numerous from Brainerd and Robinson (1951) to Ester (1981).

Data analysis techniques in seriation have been introduced by Kendall (1971) with a multidimensional scaling algorithm applied to an occurrence table modified by a special 'circular product' transformation: the Horshu method.
In parallel, the reciprocal averaging method was used in successful applications by Goldmann (1973) in Germany, Axis Wilkinson (1974) in U.K., and Legoux (1980) in France, while the theoretical similarity with the travelling salesman problem (Wilkinson, 1971), well known in Operational Research, has lead to no real development.

Correspondence analysis has been introduced due to two of its main features (Djindjian, 1976a):

- the evidence of parabolic 'Gutmann effect' for both observations (e.g. graves, layers,...) and variables (e.g. types, attributes,...) giving the seriation, and the reorganization of the table from the first axis coordinate. This figure is the analogous to the Horseshoe curve of Kendall;

- the demonstration of the identical results obtained using C.A. and reciprocal averaging method.

C.A. is used moreover as an interactive serial analysis, for the elimination of observation outliers, aberration variables, non-chronological factors, and for exposing complex seriation patterns (Djindjian 1980b, 1985a).

A toposeriation method has been developed (Djindjian, 1985a) using simultaneously C.A. and a k-means-like cluster analysis with chronological constraints on the spatial distribution of the graves of the cemetery.

Numerous studies of seriation using C.A. have been published: Leredde and Perin (1980), Boeliche, Stehli and Zimmermann (1980), Courbin (1983), Audouze (1984), Slachmuylder (1984), etc...

A special seriation package, using C.A., has been developped by Scollar and Alii (1984).

20.5 Other applications

C.A. has been used in physical and chemical characterisation and provenance studies, although the use of PCA or discriminant analysis seem to be more suitable to quantitative measures, except in the case of particulary difficult distributions where disjunctive coding is necessary.

In palaeoclimatology or in palaeoenvironnemental studies, C.A. is a good alternative for exposing the climatic factors, and for the construction of palaeotemperature curves.

Roux (1979) has proposed an improvement with C.A., of the PCA based method by Imbrie and Kipp (1971). Numerous studies have been realised in France with this method (Laurin and Rousseau, 1985), (Gasse, 1986), (Denys, 1985), etc...

C.A. has been also used to find evidence of phyletic evolution (Chaline and Laurin, 1984), and taxonomic analysis.

20.6 Data analysis as a cognitive process

The use of data analysis techniques in a cognitive process in Archaeology has been accomplished by a number of different and contradictory tendencies:

- Binford (1966), in his famous case study on mousterian assemblages, employed factor analysis, following the psychometricians, looking for the hidden factors: in this case, functional patterns.

- Benzecri (1973) used C.A., in a total inductive perspective, hypothesising de facto, the transparency of the archaeologist in the construct, to find the internal patterns in the data, from a supposedly exhaustive description.
• Doran and Hodson (1975) used data analysis as a multivariate reduction method on observations only, without trying to formalize any rules on the description (variables).

• Gardin (1979), as a result of a semiologic analysis, refused to attribute to data analysis any non-tautologic capabilities of data patterning considering that patterns give only information already contained in the data (in fact introduced by the archaeologist).

• Djindjian (1980, 1985b) used data analysis as a tool in an interactive cognitive process between intrinsic data and extrinsic data, applying a systemic approach.

The above considerations demonstrate the need to separate the three levels of the data analysis role in the archaeological constructs:

• the technical level (capabilities and limits of data analysis techniques for each case study),

• the methodological level (integration of data analysis into an archaeological method: seriation, culture pattern study, typology, ...)

• the cognitive level (explicitation of the interpretative propositions for any use of archaeological methods).

20.7 Conclusion

The data analysis techniques of the French school have been applied to quantitative Archaeology for about fifteen years. Their specific features, in comparison with other data analysis techniques, gave to French archaeologists a more powerful tool for the resolution of some of the major archaeological methodological problems of the last forty years: typology, spatial analysis, seriation, culture pattern studies, provenance studies, palaeoclimatology.

Among these techniques, Correspondence analysis plays a particular role, given by the R+Q-mode, and the non-linear features of the X² s' distance; C.A. allows a good compromise between the semiological approach of perception of the archaeological information and the mathematical processing of algorithms, in an interactive archaeologist-computer cognitive process. It is the reason why data analysis techniques, following the precursory views of Hodson (1969): 'Searching for structure within multivariate archaeological data', placed in a general theoretical context: semiologic in Gardin (1979), systemic in Djindjian (1980a), may have a resonance in the large debate about cognitive processes in Archaeology.

Bibliography


COURBIN, P. 1983. *Classement informatisé des skyphoï protocorinthiens*, X.


200


