Journal of Data Science **6**(2008), 135-154

A Statistical Approach for Dating Archaeological Contexts

L. Bellanger[1], R. Tomassone[2], P. Husi[3]
[1] *Université de Nantes,* [2] *Institut National Agronomique and*
[3] *Université de Tours*

Abstract: This paper describes a statistical model developing from Correspondence Analysis to date archaeological contexts of the city of Tours (France) and also to obtain an estimated absolute timescale. The data set used in the study is reported as a contingency table of ceramics against contexts. But, as pottery is not intrinsically a dating indicator (a date is rarely inscribed on each piece of pottery), we estimate dates of contexts from their finds, and we use coins to attest the date of assemblages. The model-based approach uses classical tools (correspondence analysis, linear regression and resampling methods) in an iterative scheme. Archaeologists may find in the paper a useful set of known statistical methods, while statisticians can learn a way to order well known techniques. No method is new, but their gathering is characteristic of this application.

Key words: Chronology, pottery, quantitative fabric analysis, regression model, resampling methods, simple correspondence analysis in archaeology.

1. Introduction

The introduction of data analysis in archaeology took place long since, see for instance updates in Buck (1999). The principal techniques of interest were first of all multidimensional scaling and cluster analysis. In the eighties multivariate analysis also consisting of Principal Component Analysis and Correspondence Analysis (hereafter CA), were developing in the European countries. These statistical techniques have made a major contribution to almost all archaeological problems like for example, chronological seriation. But they are generally viewed as exploratory tools, allowing the researcher to avoid the fastidious computations of elementary statistics, and also to summarize the data.

In this paper, we describe a statistical model developing from CA to date archaeological contexts of the city of Tours (France) and also to obtain an estimated absolute timescale. The important number of excavations achieved with the same system of data recording during the last thirty-five years (1968-2002) explains the interest of Tours (see Galinié, 2000). As pottery is a very good chronological indicator, its quantification can prove crucial, not only when comparing different archaeological contexts (or sets), but also as a help to answer certain archaeological and historical questions. The data set used can be reported as a contingency table of archaeological contexts against a particular archaeological material (pottery): rows represent different fabrics and columns specify archaeological contexts. The columns are separated into two groups. The first one includes archaeological contexts for which dates are attested by coins, the second one includes contexts whose dates are badly defined or unknown.

Several attempts have been made to assess the comparison of ceramic assemblages to establish absolute date of contexts, but never on the scale of a whole town ((Djindjian, 1991); (Tyers and Orton, 1991); (Baxter, 1994) and (Orton, 2000)) and this statistical procedure tackled in three steps:

1. Investigation of the relationship between contexts and fabrics using correspondence analysis ((Benzécri, 1973); (Greenacre, 1984) and (Moreau, Doudin and Cazes, 2000)) to obtain chronological patterns.
2. Use of the secure representation of the contexts obtained previously, for the purpose of estimating their date with a **regression model.**
3. Model checking as an essential component of this fitting process, including **resampling methods** (jackknife and bootstrap).

This method provides an effective complementary tool for dating archaeological contexts.

2. Archaeological Questions and Data Corpus

2.1 Data corpus

An important part of the archaeological study of pottery is the comparison of ceramic assemblages in terms of their compositions. A ceramic assemblage can be characterized by the proportions of different fabrics of pottery with which it is made up. Various measures for quantifying pottery may be used, here we choose *minimum vessels count* in which fragments are assumed to belong to the same vessel unless they can be shown to belong to different ones.

Finally, we use a $(r \times c)$ data matrix called N consisting in:

- The 49 sets as columns $(r = 49)$; the sets are ceramic assemblages, representing the different human occupation stages of excavation (building, possession and destruction of an edifice). They come from 6 excavations in the city of Tours. The use of an homogeneous and precise stratigraphic recording system allows us to obtain many sets distributed in long sequences .

- The 186 fabrics of pottery that represent the types of pottery in ceramic assemblages, as rows ($c = 186$). Figure **??** presents four studied fabrics.

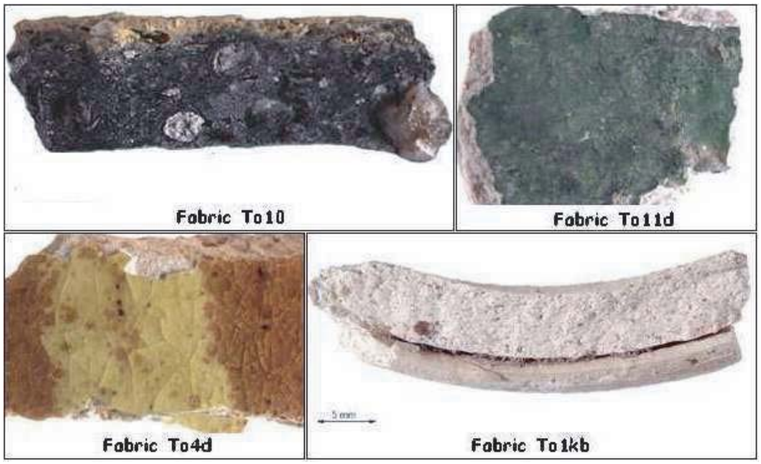


Figure 1: Four studied fabrics

2.2 Absolute dating ceramic assemblages in Tours

As pottery is not intrinsically a dating indicator (a date is rarely inscribed on each piece of pottery), we estimate dates from contexts of their finds, and we use coins to attest the date of assemblages. A more secured process to obtain more reliable dates consists in conserving only the coin dates associated with a building which is itself reliably dated by documentary evidence, and eliminating or giving less importance to isolated coins. In our case, we retain 21 sets out of 49 for which absolute date is attested by coins, the others have no absolute date (28).

A statistical approach, as we shall see in the following sections, may point to dating accuracy of coins: an important chronological time-lag between dating of coins and the estimated absolute dating of ceramic assemblages, may lead to some **archaeological reinterpretations of stratigraphic sequences** (see Figure 2).

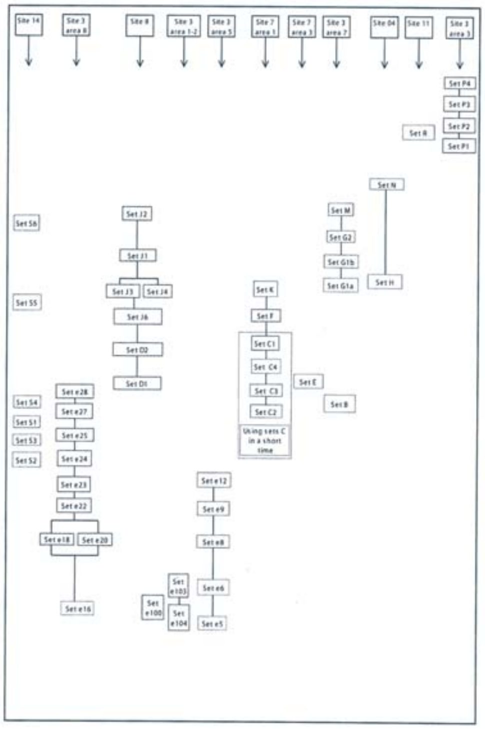


Figure 2: Studied sets stratigraphical sequences

3. Statistical Procedure

The statistical procedure uses classical tools (correspondence analysis, linear regression and resampling methods); its originality is the use of an iterative scheme of the three steps as explained above in paragraph 1.

3.1 Correspondence Analysis

CA is well known as a technique for the display of rows (in our case fabrics) and columns (in our case sets) of a two-way contingency table as points in a low-dimensional vector space that is readily interpretable when displayed graphically.

When working on the $(r \times c)$ data matrix N, the best known method (and among the oldest and most widely used one of multivariate techniques) is *principal component analysis* (PCA) which may be presented in different ways. It involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. It is a linear transformation that chooses a new coordinate system for the data set as such that the greatest variance by any projection of the data set comes to lie on the first axis (called the first principal component), the second greatest variance on the second axis, and so on. The objectives of principal component analysis are to discover or to reduce the dimensionality of the data set and to identify new meaningful underlying variables.

PCA is a method pertaining to *factor analysis family*; *CA* is another member of this family: it simultaneously characterizes the relationship among the rows and also among the columns of a data matrix preferably, but not only, as a contingency table. It may be outlined as a technique that can be used to study interaction in a two-dimensional contingency table. The *CA* coordinates are analogous to those derived from a *PCA*, except that they are derived by partionning the total chi-square statistic for the table, rather than total variance. In this section, we will briefly examine correspondence analysis, but a full account is available for example in (Greenacre, 1984).

If we write the table $N = [n_{ij}], i = 1, \cdots, r; j = 1, \cdots, c$, we first convert it to a new table of observed proportions:

$$\mathbf{P} = [p_{ij} = \frac{n_{ij}}{n_{..}}] \in \mathcal{M}_{r\times c}, \text{where } n_{..} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}.$$

The row and column marginal frequencies are given by the vectors:

$$\mathbf{p}_{i\cdot} = [\frac{n_{i.}}{n_{..}}] \text{ and } \mathbf{p}_{\cdot j} = [\frac{n_{.j}}{n_{..}}].$$

Diagonal matrices constructed from these vectors are denoted by $\mathbf{D}_r \in \mathcal{M}_{r\times r}$ and $\mathbf{D}_c \in \mathcal{M}_{c\times c}$ respectively.

We may define:

- For each row i, a row profile $\mathbf{r}_i = [\frac{n_{ij}}{n_{i.}}]$ (a $(c \times 1)$ vector which is a row conditional proportions);

- For each column j, a column profile $\mathbf{c}_j = [\frac{n_{ij}}{n_{.j}}]$ (a $(r \times 1)$ vector is a column conditional proportions);

- The weighted average of the r profiles can be used to obtain an *average row profile* by $\sum_{i=1}^{r} \mathbf{r}_i(n_{i.}/n_{..}) = \mathbf{p}_{\cdot j}$ which is the $(c \times 1)$ vector of column frequencies;

- The weighted average of the c profiles can be used to obtain an *average column profile* given by $\sum_{j=1}^{c} \mathbf{c}_j(n_{.j}/n_{..}) = \mathbf{p}_{i\cdot}$ which is the $(r \times 1)$ vector of row frequencies.

When the two variables are independent, the frequencies to be expected in the contingency table should not differ a lot of:

$$e_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \Longleftrightarrow p_{ij} = p_{i\cdot} p_{\cdot j}.$$

Having these notations in mind, we can analyze the matrix difference $\mathbf{P} - \mathbf{p}_{i\cdot}\mathbf{p}_{\cdot j}^{'}$ which is a measure of the deviation from independence.

For this matrix the *Singular Value Decomposition* subject to the conditions $\mathbf{A}^{'}\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{I}_r$, $\mathbf{B}^{'}\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}_c$ is given by:

$$\mathbf{P} - \mathbf{p}_i\mathbf{p}_j^{'} = \mathbf{A}\mathbf{D}_\mu\mathbf{B}^{'} = \sum_{k=1}^{K} \mu_k \mathbf{a}_k \mathbf{b}_k^{'}$$

where $\mathbf{D}_r^{-1} = diag(n_{i\cdot}^{-1})_{i=1,\ldots,r}$, $\mathbf{D}_c^{-1} = diag(n_{\cdot j}^{-1})_{i=1,\ldots,c}$, $K = \min[(r-1),(c-1)] = rank\left[\mathbf{P} - \mathbf{p}_i\mathbf{p}_j^{'}\right]$, the K columns of $\mathbf{A} \in \mathcal{M}_{r\times K}$ and $\mathbf{B} \in \mathcal{M}_{c\times K}$ are denoted by $\mathbf{a}_k \in \mathbb{R}^r$and $\mathbf{b}_k \in \mathbb{R}^c$ respectively and μ_k are the diagonal elements of the diagonal matrix $\mathbf{D}_\mu$.

Using $\mathbf{A}$ and $\mathbf{B}$ we may obtain coordinates for the row and column profile deviations; $\mu_k(\leq 1)$ is a measure of the intensity of relation between rows and columns. For example, a two-dimensional representation the column coordinates, also named *correspondence map of the column points*, may be represented as $b_{jk}, j = 1, 2, \ldots, c; k = 1, 2$.

We call *total inertia or inertia* the sum of the squares of the singular values:

$$tr[\mathbf{D}_r^{-1}(\mathbf{P} - \mathbf{p}_{i\cdot}\mathbf{p}_{\cdot j}^{'})\mathbf{D}_c^{-1}(\mathbf{P} - \mathbf{p}_{i\cdot}\mathbf{p}_{\cdot j}^{'})] = \sum_{k=1}^{K} \mu_k^2 = G^2/n_{\cdot\cdot}$$

Total inertia can be viewed as a measure of the magnitude of the total row squared deviations or equivalently the magnitude of the total column squared deviations and the *SVD* can be used to allocate the total inertia to various dimensions and reflects the spread of points around the centroid. Total inertia is linked with the so-called Pearson Chi-square statistic G^2 used for testing the null hypothesis H_0 *independence between rows and columns* of a contingency table. It may be written in different ways, the classical one being:

$$G^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \underset{(H_0)}{\approx} \chi^2\left((r-1)(c-1)\right)$$

It may be easily interpreted as a global measure of distances over rows or over columns and is known as the Chi-squared metric. It may also be written:

$$G^2 = \sum_{i=1}^{r} n_{i.}(\mathbf{r}_i - \mathbf{p}_{i\cdot})' \mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{p}_{i\cdot}) = \sum_{j=1}^{c} n_{.j}(\mathbf{c}_j - \mathbf{p}_{\cdot j})' \mathbf{D}_r^{-1}(\mathbf{c}_j - \mathbf{p}_{\cdot j})$$

Total inertia may be interpreted as the percentage of inertia(variance) in the original correspondence table explained by all the computed dimensions in the correspondence analysis. However, usually only the first dimensions are used in the correspondence map, so the effective model will explain a percentage of inertia in the original table equal to the sum of eigenvalues for these first dimensions only. The adequacy of a one-or two-dimensional representation of the residuals is judged by the proportion of the inertia explained by each dimension.

CA has a special interest due to its useful properties:

- With the Chi-squared metric, if two row profiles are identical they may be replaced by a single row profile that is the sum of the two profiles. This collapsing of the two rows will not affect the geometry of the column profiles. If two row profiles are identical they occupy identical positions in the row space. A symmetrical result is for columns.

- The relations (transition formulaes) between the row coordinates and the column coordinates (with some precaution) allow to display both row and column points in a single map, more or less as the classical *biplot*.

$$\begin{aligned}
\mathbf{a}_k &= \frac{1}{\mu_k}\mathbf{D}_r^{-1} N \mathbf{b}_k \text{ so that } a_{ik} = \frac{1}{\mu_k}\sum_{j=1}^{c}\frac{n_{ij}}{n_{i\cdot}} b_{jk} \\
\mathbf{b}_k &= \frac{1}{\mu_k}\mathbf{D}_c^{-1} N \mathbf{a}_k \text{ so that } b_{jk} = \frac{1}{\mu_k}\sum_{ji1}^{r}\frac{n_{ij}}{n_{\cdot j}} a_{ik}
\end{aligned}$$

- The method is a special case of *canonical correlation*, where one set of entities (categories rather than variables as in conventional canonical correlation) is related to another set.

- The extension of results for a two-way contingency table is easily done for a multi-way one.

All these elements show the interest of *CA* as a good substitute of *PCA* in a lot of practical situations and not only for contingency tables; an important literature exists on the topic, mainly in French.

It was first applied to archaeology in France in 1975, following the publication of Benzécri (Benzécri, 1973), see for example (Djindjian, 1980) a first synthesis of different applications to archaeology). In our study, it appears to be the obvious technique because we are seriating archaeological contexts in terms of counts of the types of pottery they contain. But it is important to underline that contexts containing pottery assemblages could be some distance apart from another, may have been used for different purposes (e.g. cooking and eating) or there may be time-lags in the way of life changes between one context and the next. Then, they generally cannot be put in a strict sequence, (Laxton, 1990) has identified a potential problem for the use of *CA* when the standard pattern of first appearance, a rise in population, a peak, a decline or a disappearance is not satisfied. In that case, *CA* is not guaranteed to produce a correct sequence even when the types are chronologically diagnosed ((Baxter, 1994), pages 118-123). As developed in section 2, we see here too the importance of the archaeological choice of the studied data corpus to obtain consistent statistical results. Indeed, it will be important to justify the association of the order in the data (if such an association exists) with time rather than some other dimension.

The *CA* on the N matrix (186×49 at the beginning of the iterative scheme) provided the classical results: eigen values, column and row factors coordinates and others as contributions.

3.2 Regression on factors

Assuming y_j represents the value of the response variable (coin date) on the jh individual (a dated set), and that $b_{j1}, b_{j2}, \dots, b_{jp}$ represent the individual's values on p explanatory variables (the first $p(=4)$ column coordinates of the *CA*), with $j = 1, \dots, n(= 21)$. The multiple linear regression model is given by:

$$y_j = \beta_0 + \beta_1 b_{j1} + \ldots + \beta_p b_{jp} + \varepsilon_j; \qquad j = 1, \dots, n$$

where ε_j ,$j = 1, \dots, n$ are assumed to be independent random variables having a Normal distribution $\mathcal{N}\left(0, \sigma^2\right)$; and $\beta = [\beta_0, \beta_1, \dots, \beta_p]'$ is the parameter vector of the model, known as the *regression coefficients.*

The multiple regression model can be written most conveniently using matrices and vectors as:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_n]'$, $\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & b_{11} & b_{12} & \ldots & b_{1p} \\ 1 & b_{21} & b_{22} & \ldots & b_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & b_{n1} & b_{n2} & \ldots & b_{np} \end{bmatrix} = [1_n, \mathbf{b}_1^{(1)}, \mathbf{b}_2^{(1)}, \ldots, \mathbf{b}_p]$$

and vector $\mathbf{b}_k^{(1)}$ contains the k column coordinates of the n dated sets derived from CA.

Assuming that $\mathbf{X}'\mathbf{X}$ is nonsingular, then the least-square estimator of the parameter vector β is:

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}$$

The regression analysis can be assessed using the classical analysis of variance table.

The statistical procedure is summarized in Appendix.

3.3 Resampling methods (*CA* and date expected value)

There are many texts and articles on the subject; interested readers can find details on resampling methods in a number of text books, such as (Davidson, 1997) which also give many examples and practicals. We compare the results of different resampling methods.

In the following, each row in $\mathbf{X}$ is noted $\mathbf{x}_j = [1, b_{j1}, b_{j2}, \ldots, b_{jp}]$. It represents the values of the explanatory variables for one of the individuals (sets) in the sample, with the addition of unity to take account of the parameter β_0.

The pottery assemblages variability: data corpus resampling

- *Jackknife*: to see if fabrics may have a strong influence on date estimation, we have first used the jackknife on the N matrix. In our jackknife procedure, the given statistic is recalculated for $S = 186$ fixed data sets that are subsets of the original one. The following algorithm summarizes the calculation of estimated dates and associated confidence sets:

  Algorithm 1. For $k = 1, \ldots, S$

  (a) let N_{-k} be the N matrix minus the kth row; then

  (b) calculate the factors coordinates of the columns (sets) of N_{-k} matrix using a CA;

  (c) fit least squares regression to $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_j$ is the p-vector of the selected factor coordinates of columns of N_{-k}, giving estimates $\widehat{\beta}^{(k)}_{(Jack)}$ for β.

We deduce a confidence interval for y_i.

- *Bootstrap*: in the bootstrap procedure the recomputations are based on B bootstrap data sets randomly chosen (bootstrap samples) of size 186 × 49 generated from the original one with B chosen so that the summary measure of the individual statistics is nearly good when taking $B = \infty$. In our case, B was fixed to 1000 in all bootstrap applications.

  Algorithm 2. For $b = 1, \ldots, B$

  (a) generate a replicate population $N^{(b)}$ by sampling 186 times with replacement from the rows of N;

  (b) calculate the factor coordinates of the columns (sets) of $N^{(b)}$ matrix using a CA;

  (c) fit least squares regression to $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_j$ is the p-vector of the selected factor coordinates of columns of $N^{(b)}$, giving estimates $\widehat{\beta}^*_{(b)}$ and $s_{i\,(b)}^{2^*}$ for the residual mean square σ^2.

  We obtain a bootstrap percentile confidence interval for y_j. The method just presented for obtaining a nonparametric confidence interval for the expected value of date is the *bootstrap percentile* one. It is the simplest but not necessarily the best performing bootstrap method.

The dates variability: model residual resampling

- As we suppose that the $\mathbf{x}_j$ are non-random in our linear regression model, we decided to use the resampling scheme named *bootstrapping residuals* or *bootstrap based on residuals* proposed by Efron (see for example (Efron and Tibshirani, 1993)). The following algorithm summarizes the calculation of estimated dates and confidence sets associated:

  Algorithm 3.

  (a) calculate the factor coordinates of the columns (sets) of N matrix;

  (b) fit least squares regression to $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$: our model can be identified as (β, F_ε), where F_ε is the common unknown distribution of random variables ε_i centred and independent identically distributed, abbreviated to i.i.d. The parameter β is estimated by the least square estimator $\widehat{\beta}$ and F_ε by the empirical distribution $\widehat{F_\varepsilon}$ putting mass n^{-1} to r_j, j= $1, \ldots, n$, where $r_j = y_j - \mathbf{x}_j\widehat{\beta}$ are the residuals;

  (c) model-based resampling:

  For $br = 1, \ldots, B$:

(i) for $j = 1, \ldots, n$
- randomly sample: generate i.i.d. data $\varepsilon_1^*, \ldots, \varepsilon_n^*$ from $\widehat{F}_\varepsilon$ (that is from $r_j, j = 1, \ldots, n$); then
- define $y_j^* = \mathbf{x}_j \widehat{\beta} + \varepsilon_j^*$.

(ii) fit least squares regression to $(\mathbf{x}_1, y_1^*), \ldots, (\mathbf{x}_n, y_n^*)$, giving estimates $\widehat{\beta}^*_{(br)}$ and $s^{2^*}_{(br)}$ for the residual mean square σ^2;

(d) calculate the bootstrap percentile confidence interval for y_j.

4. Results[1]

4.1 Correspondence analysis

The proportion of inertia accounted for by dimensions 1 to 7 are presented in Table 1. The first seven dimensions account for 64.1% of the inertia and therefore give an accurate representation of the fabrics/sets relationship.

Table 1: Inertia percentage accounted for by the first seven dimensions of CA.

Factor	1	2	3	4	5	6	7	Sum
%	24.2	11.6	8.7	6.7	4.9	4.2	3.7	64.1

We have analyzed in details:

- The *contribution of set points to dimensions* to intuit the meaning of correspondence dimensions. The contribution of points to dimensions shows the percentage in order of inertia of a particular dimension which is explained by a point. By looking at the more heavily loaded points, one may deduce the meaning of a dimension. Contribution of points to dimensions will add up to 1.0 across the categories of any one variable. We demoted sets points $\{e27, e28, P1, N, R\}$ from being an active one (category values of the sets variables used to compute the dimensions) to being a supplementary one because they tend to influence the definition of dimensions. Then we reintroduce these sets variables by computing their new coordinates as *supplementary* ones. These can be plotted on the correspondence map also. In our case we use supplementary sets points to handle outliers that may unduly affect the computation of dimensions.

[1] Computations were carried out using the S-Plus or R statistical softwares

- *Contributions of dimensions* (also known as the *quality of representation* of the description of a point) to sets (column points). These reflect how effective the correspondence analysis model is in explaining any given point. That is, the contribution of dimensions to points is the percentage of variance in a point explained by a given dimension. One would like the points on which one's analysis focuses to have a high contribution of dimensions to points value. Less analytic focus must be placed on points which are not well described by the model. The sum of contributions of dimensions to points will add up to 1.0 across all dimensions for a given point in the full solution where all possible dimensions are computed.

A plot of the first two coordinates and also coordinates 1 and 3 are shown in Figure 3:

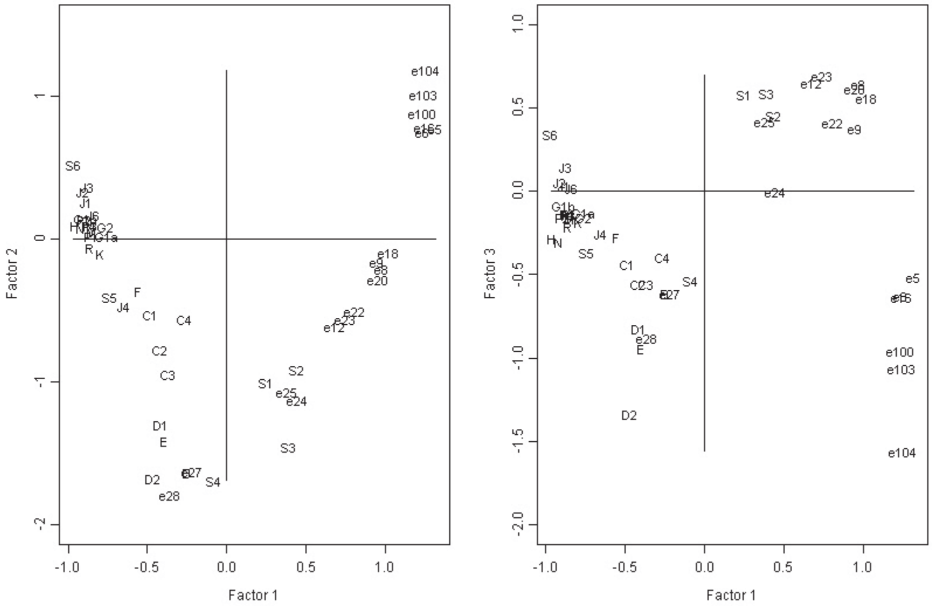


Figure 3: Correspondence Analysis (e27, e28, P1, N and R as supplementary sets)

The interpretations of this diagram are the following:

- It is well known that a chronological sequence should be represented by a 'horse-shoe' shaped curve on a *CA* plot (also known as the 'arch' or 'Guttman' effect). It arises for reasons discussed by Greenacre ((Greenacre, 1984), section 8.3). Indeed, the first factor is clearly linked with a trend that we shall identify as the best chronological seriation order because the types in terms of which the units are being described are chronologically sensitive (e.g. independent stratigraphic evidence). Since the second (resp. third) factor is parabolic (resp. cubic), the regression model looks like a polynomial one; but when we have tried to use a polynomial model with $\mathbf{b}_1^{(1)}, \left(\mathbf{b}_1^{(1)}\right)^2 \ldots$ the results were not improved and are in fact even worse.

- We have more or less 3 groups of sets: {e100, e103, e104, e16, e6, e5} on the right are the oldest, {S1, e25, e24, S3, S2, e23, e22, e20, e18, e12, e9, e8} are in an intermediate position, and all the others are the more recent ones. This pattern might reflect social and functional differences within the city through time. It would be interesting to introduce the forms and not only the fabrics (e.g. cooking-pots, jugs, bowls,..) to have perhaps a clearer indication of the interpretation of this pattern; doing so, we would study typological criteria (fabrics and forms) with regard not only to chronological questions but to functional aspects of pottery.

4.2 Regression

We have dates attested by coins for a limited number sets (21) as shown in Table 2. Sometimes, we have different available coins; it was decided, on archaeological grounds, to choose one of them. Nevertheless we must remember this choice during interpretation and data validation, to perhaps modify our first choice.

Table 2: Values of observed dates for dated sets

set	e16	e18	e20	e22	e8	e9	e27
date	850	1100	1100	1100	1100	1100	1296
set	e28	D1	E	D2	F	G1a	J2
date	1316	1350	1350	1436	1461	1470	1476
set	G1b	J3	H	J1	M	R	P1
date	1488	1488	1510	1540	1540	1625	1640

As we are interested in prediction, associated with prediction intervals, for $n = 21$ dated and 28 undated sets, it is firstly necessary to select a subset of regressors (in our case, subset of the first p CA column coordinates). Hopefully, the first ones were always the best in this selection; in some cases we could have chosen fewer but not in their natural order of importance. This option would

Table 3: Selected model

Corpus	Selected model	s	df	R^2
N	b_1, b_2, b_3, b_4	40.70	16	0.9709

s: residual standard deviation; df: degrees of freedom; R^2: determination coefficient.

Table 4: Predicted dates (90% confidence intervals)

set	date	Lower	Pred	Upper	set	date	Lower	Pred	Upper
e16	850	807	875	943	e6	?	813	880	947
e18	1100	1037	1069	1101	e23	?	1128	1166	1205
e20	1100	1066	1100	1133	e12	?	1157	1195	1233
e22	1100	1097	1127	1157	e24	?	1182	1212	1242
e8	1100	1061	1094	1128	S2	?	1209	1246	1284
e9	1100	1051	1080	1108	e25	?	1221	1261	1301
$\underline{\text{e27}}$	1296	1307	1339	1371	S3	?	1235	1290	1345
$\underline{\text{e28}}$	1316	1324	1359	1393	S1	?	1262	1306	1351
D1	1350	1336	1363	1390	S4	?	1285	1318	1352
E	1350	1333	1362	1392	C4	?	1327	1344	1361
$\underline{\text{D2}}$	1436	1310	1349	1388	B	?	1343	1376	1408
$\underline{\text{F}}$	1461	1398	1417	1436	C3	?	1356	1376	1397
G1a	1470	1461	1483	1504	C2	?	1369	1388	1407
J2	1476	1465	1498	1532	C1	?	1371	1390	1409
$\underline{\text{G1b}}$	1488	1497	1521	1545	S5	?	1422	1446	1470
J3	1488	1461	1495	1530	J4	?	1432	1451	1471
$\underline{\text{H}}$	1510	1525	1549	1574	S6	?	1449	1498	1547
$\underline{\text{J1}}$	1540	1476	1506	1535	K	?	1455	1479	1502
M	1540	1505	1527	1549	J6	?	1461	1490	1519
R	1625	1592	1652	1713	G2	?	1475	1496	1518
$\underline{\text{P1}}$	1640	1541	1571	1601	P2	?	1540	1569	1598
e104	?	669	780	892	P3	?	1550	1586	1622
e103	?	742	831	919	P4	?	1550	1585	1619
e100	?	766	848	930	N	?	1571	1617	1663
e5	?	802	868	933					

Underlined values (lower or upper) indicate that the true value is out of range.
For dated sets the residuals range is $-43 : 87$.

have introduced more difficulties later, as we shall explain when using resampling techniques. Individual regression coefficients were assessed by using the ratio $\widehat{\beta}_j / SE\left(\widehat{\beta}_j\right)$: the four regression coefficients were significant[2], the residual standard deviation was minimal; so this model (Table 3) seems to be a good candidate for a predictive purpose.

[2] The null hypothesis H_0 : "$\beta_j = 0$" of t-test on the corresponding individual regression coefficients was rejected ($j = 1, ..., 4$).

This model is coherent with archaeological arguments, it will play the role of our reference model. We may easily compute the predicted dates and intervals for each quantification (Table 4).

Their precision will be scrutinized in section 4.3. But at first glance, we may look at the mean range of different 90% confidence intervals. Their values are 75 years. The residuals distribution is based on a too limited set (21 residuals) to validate or invalidate the Gaussian error distribution implied by our regression model (Figure 4).

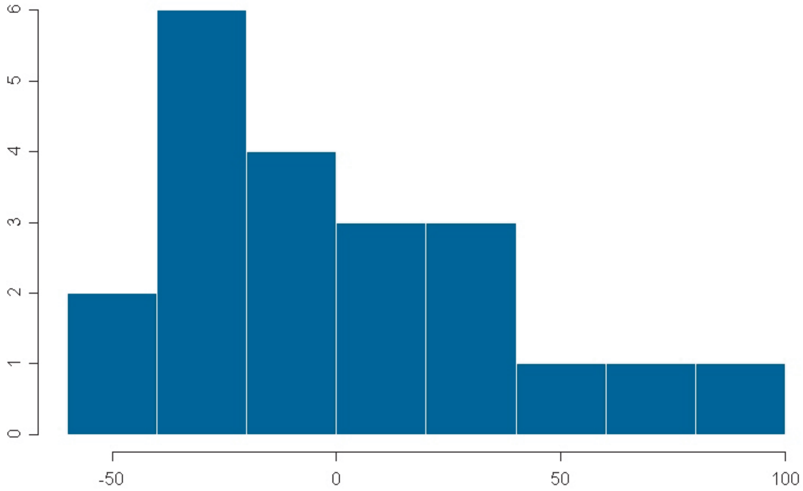
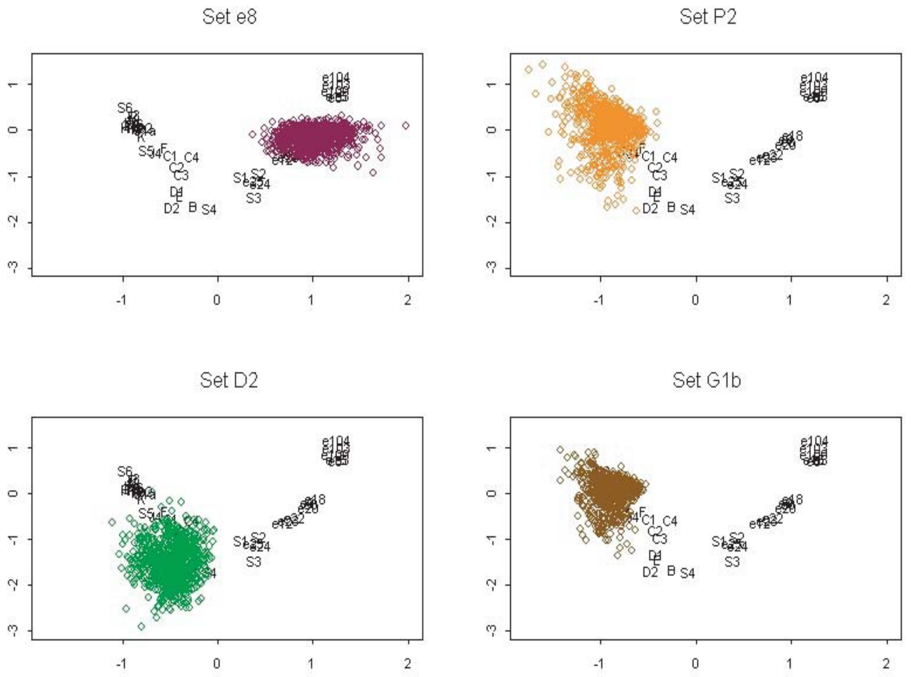


Figure 4: Histogram of residuals

From an archaeological point of view, the dates of sets, estimated by means of pottery, are quite correct. The chronological bracketing of sets within a half-century time span (90% predictive interval) is very precise for an archaeologist. In fact, as often in applied statistics, the badly fitted data are rather more interesting to analyze, because they oblige the archaeologist, being driven into a corner, to scrutinize his data more deeply (for more details see (Husi *et al.*, 2000), (Husi *et al.*, 2006) and (Bellanger *et al.*, 2006).

Since we are interested in date prediction, in the following section, we construct confidence sets for dating by using resampling methods. These computer-intensive methods are useful when inference is based on a complex procedure for which theoretical results are unavailable or difficult to use. Indeed in our case, nonparametric resampling can take into account the two main sources of errors:


Set e8

Set P2

Set D2

Set G1b

Figure 5: Data clouds of four selected sets

- The pottery assemblages variability;
- The dates variability.

4.3 Resampling

Before having archaeological interpretation of both results, we may note the fundamental differences between the two schemes:

- Under pottery *data corpus resampling* (in abbreviate *resampling cases*), the design matrix N may change for each one and is generally not equal to the original one. Of course, each factor differs to a certain extent (it is the reason why we have chosen a fixed number of factors for quantification, taking as an assumption that the matrix design for regression is defined in a fixed subspace of dimension 4). This means that in our case with quite a large data set, the design matrix is quite stable except if some few influential fabrics exist. The framework of results interpretation will be similar to the jackknife one. This scheme also allows us to assess the statistical stability of the *CA* scatterplots (see (Greenacre, 1984); (Lebart, Morineau and Piron, 1995) and (Ringrose, 1992) for application to archaeological data). It is more robust with regards to uncertainty of pottery data. Because performing a new *CA* for each replicate matrix would lead to all the points' coordinates being relative to different axes, we simply convert bootstrapped sets of column profiles into points on the *CA* coordinate system calculated from the original data. Moreover, as in practice it is only possible to examine a small number of clouds of points on one plot, we presented only the 'confidence cloud' of 4 selected sets (Figure 5).

- Under model *residuals resampling*, we can say that the scheme is more efficient if the model is correct. This assumption is difficult to ascertain; the results will point out the archaeological sets with great influence using a jackknife after bootstrap inquiry. We may imagine that the results will be more precise if we admit that uncertainty doesn't concern data themselves.

When looking at the results (Table 5 for fitted dates), we can see that they are generally homogenous; but what is more interesting is to note **some discrepancies** in some of them which oblige us to try to interpret the reason why is happened. To give two examples: e20, its jackknife estimation is quite different, when other estimations are good; D2: always underestimated. This means, once again, that a statistical analysis gives results, but is always **a tool that facilitates further inquiry**.

Table 5: Fitted dates and residuals for N quantifications. P: classical model, J: jackknife estimation, Bd: resampling cases.

		Prediction			Residual				Prediction			Residual			
set	date	P	J	Bd	rP	rJ	rBd	set	date	P	J	Bd	rP	rJ	rBd
e16	850	875	875	878	-25	-25	-28	F	1461	1417	1421	1419	44	40	42
e18	1100	1069	1060	1070	31	40	30	G1a	1470	1483	1457	1483	-13	13	-13
e20	1100	1100	1145	1102	0	-45	-2	J2	1476	1498	1504	1501	-22	-28	-25
e22	1100	1127	1141	1128	-27	-41	-28	G1b	1488	1521	1509	1524	-33	-21	-36
e8	1100	1094	1070	1091	6	30	9	J3	1488	1495	1500	1500	-7	-12	-12
e9	1100	1080	1052	1081	20	48	19	H	1510	1549	1561	1546	-39	-51	-36
e27	1296	1339	1338	1329	-43	-42	-33	J1	1540	1506	1505	1510	34	35	30
e28	1316	1359	1351	1351	-43	-35	-35	M	1540	1527	1522	1530	13	18	10
D1	1350	1363	1359	1373	-13	-9	-23	R	1625	1652	1697	1615	-27	-72	10
E	1350	1362	1357	1369	-12	-7	-19	P1	1640	1571	1569	1567	69	71	73
D2	1436	1349	1344	1368	87	92	68								

Another kind of information is provided by the estimated distribution of date given by both bootstraps. The differences between estimation are in a short interval ($-10 : +10$), except for some specific sets. The minimal and maximal values have no intrinsic interest for resampling cases; but they may indicate **sensitivity** to some fabrics deleted for one of 1000 samples.

5. Conclusion

Archaeology often treats multivariate analysis as a mechanical means to justify some archaeological reasoning and it is often used as such. It is important to emphasize that it would be naive to presume that a mechanical process can by itself act as a substitute for archaeological and statistical knowledge. The way of thinking in terms of research process is important here. *'No matter how sophisticated methods are, or may become, they will never be able to make a judgment of relevance between the individual variables. A [archaeological] judgment of relevance has to be made before analysis starts, it has to continue throughout the analyses, and it is entirely the responsibility of the archaeologist.'* (Madsen, 1988), page10).

In this article we used model-based statistical methods to establish absolute date of archaeological contexts and an archaeological chronology of the city of Tours. But we do not use prior information about the relative chronological order of some of the contexts that is available. The use of Bayesian statistical approaches might be seen as an appealing feature.

Appendix

The statistical procedure is summarized as follows:

We write $N = [N^{(1)} : N^{(2)}] \in \mathcal{M}_{186\times 49}$. The first 21 columns of N are for the dated sets $y^{(1)}$; while the last 28 columns of N are for the undated sets $y^{(2)}$. The rows of N represent the 186 fabrics under consideration. For correspondence analysis (CA), N is reduced to a 4×49 matrix of the form $B = [B^{(1)} : B^{(2)}]$, where $B^{(1)} \in \mathcal{M}_{4\times 21}$ and $B^{(2)} \in \mathcal{M}_{4\times 28}$. The 4 rows of B, each of dimension 1×49, are $\mathbf{b}_j, j = 1, 2, 3, 4$. We also decompose $\mathbf{b}_j$ into $\mathbf{b}_j = (\mathbf{b}_j^{(1)}, \mathbf{b}_j^{(2)})$ with proper dimensions 21 and 28 respectively.

These are the first 4 CA column (sets) coordinates. Regression of $y^{(1)}$ is performed on $\mathbf{1}, \mathbf{b}_1^{(1)}, \ldots, \mathbf{b}_4^{(1)}$, with design matrix $\mathbf{X} = [\mathbf{1}, B^{(1)}]$. The predicted value of $y^{(2)}$ is based on $\mathbf{1}, \mathbf{b}_1^{(2)}, \ldots, \mathbf{b}_4^{(2)}$ with design matrix of the form $[\mathbf{1}, b^{(2)}]'$.

References

Baxter, M.J. (1994). *Exploratory Multivariate Analysis in Archaeology.* Edinburgh University Press.

Bellanger L., Husi P. Tomassone R. (2006). Une approche statistique pour la datation de contextes archéologiques. *Rev. Statistique Appliquée* **LIV**, 65-81.

Benzécri, J.-P. (1973). *Analyse des Données. Tome II: Analyse des Correspondances.* Dunod.

Buck, C. E. (1999). Archaeology in statistics. In *Encyclopedia of Statistical Science, Update Volume 3* (Edited by Kotz, S. and Johnson, N. L.), 6-11. Wiley.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

Djindjian, F. and Leredde, H. (1980). Traitement automatique des données en archéologie. *Les dossiers de l'archéologie* **42**, 52-69.

Djindjian, F. (1991). *Méthodes pour l'archéologie.* Armand Colin.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

Galinié, H. (2000). Ville, espace urbain et archéologie, col. Sciences de la ville, n°16, Maison des Sciences de la Ville, de l'Urbanisme et des Paysages, CNRS-UMS 1835, Université de Tours.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press.

Husi, P., Tomassone, R. in collaboration with Chareille, P. (2000). Céramique et chronologie : de l'analyse factorielle au modèle linéaire, Application aux sites d'habitats de Tours, *Histoire & Mesure* **XV-1/2**, 3-32.

Husi P., Bellanger L., Tomassone R. (2006). Statistical aspects of pottery quantification for dating some archaeological contexts. *Archaeometry* **48**, 169-183.

Laxton, R. R. (1990). Methods of chronological ordering. In *New Tools from Mathematical Archaeology* (Edited by Voorrips A. and Ottaway B.), 37-44. Warsaw: Scientific Information Centre of the Polish Academy of Sciences.

Lebart, L., Morineau, A., Piron, M. (1995). *Statistique exploratoire multidimensionnelle.* Dunod.

Madsen, T. (1988). Multivariate statistics and archaeology. In *Multivariate Archaeology* (Edited by Madsen T.), 7-27. Aarhus University Press.

Moreau, J., Doudin, P.-A., Cazes, P. (2000). *L'Analyse des correspondances et les techniques connexes.* Springer.

Orton, C. R. (2000). *Sampling in Archaeology.* Cambridge University Press.

Ringrose, T.J. (1992). Bootstrapping and Correspondence Analysis in Archaeology, *Journal of Archaeological Science* **19**, 615-629.

Tyers, P. A. and Orton, C. R. (1991). Statistical analysis of ceramic assemblages. In *Computer Applications and Quantitative methods in Archaeology* (Edited by Lockyear K. And Rahtz S.), 117-120. BAR International Series 565, Oxford.

Received February 1, 2005; accepted March 25, 2007.

L. Bellanger
Université de Nantes
Laboratoire de Mathé
matiques Jean Leray - UMR 6629
BP 92208, 44322 Nantes Cedex 03, France
lise.bellanger@univ-nantes.fr

R. Tomassone
Institut National Agronomique
Département de Mathématiques
75231 Paris Cedex 05, France
rr.tomassone@wanadoo.fr

P. Husi
Université de Tours
UMR 6173 CITERES
Laboratoire Archéologie et Territoires
BP 60449, 37204 Tours Cedex 03, France
philippe.husi@univ-tours.fr