CHAPTER 7

CORRESPONDENCE ANALYSIS OF ARCHEOLOGICAL ABUNDANCE MATRICES

ABSTRACT. In this chapter we discuss the Correspondence Analysis (CA) techniques used in other chapters of this book. CA is presented as a multivariate exploratory technique, as a proximity analysis technique based on Benzécri distances, as a technique to decompose the total chi-square of frequency matrices, and as a least squares method to fit association or ordination models.

Date: January 2, 2008.

CONTENTS

List of Tables

List of Figures

1. Introduction

Correspondence Analysis (CA from now on) is a technique to analyze data matrices of non-negative numbers. CA is related to *principal component analysis (PCA)* and *multidimensional scaling (MDS)*, i.e. it is a form of *proximity analysis*. CA is most frequently applied to rectangular tables of frequencies, also known as *cross tables* or *contingency tables*, alhough applications to binary incidence or presence-absence matrices are also quite common.

The most often used statistical technique for analyzing cross tables computes and tests some measure of *independence* or *homogeneity*, such as Chi-square. In the analysis of independence we investigate if the body of the table is the product of the marginals. Or, if one prefers an asymmetric formulation, if the rows of the table differ only because they have different row-totals (and the columns only differ because they have different column-totals).

Pearson's Chi-square and related measures quantify how different the observed table is from the expected table, computed from the row and column totals. Pearson-residuals are used to investigate deviations from independence. CA supplements this classical

Chi-square analysis, because it makes both a *decomposition* and a *graphical representation* of the deviations from independence.

1.1. **History.** CA has a complicated history, both in statistics and in archeology. The prehistory of CA, starting with work by Pearson around 1900 and ending with the reinvention of the technique by Fisher and Guttman around 1940, is discussed by De Leeuw [1983]. Subsequently the technique was re-reinvented under many different names, in many different countries, and in many scientific disciplines. New reincarnations still continue to appear, although at a slower pace than before, in the data mining and data analysis literature. Beh [2004] is a recent comprehensive bibliographic review.

The history of CA in archeology is discussed by Baxter [1994, p. 133-139]. Although there were some earlier applications to archeological examples in the CA literature, the credit for the introduction of the technique to archeologists usually goes to Bølviken et al. [1982]. Early applications almost without exception came from archeologists in continental Europe, under the influence, no doubt, of the French *Analyse des Données* school, under the leadership of

Benzécri [1973a; 1973b]. A good overview of these various continental archeological applications of CA is, for example, Müller and Zimmerman [1997].

It is clear from Baxter's discussion that archeologists in continental Europe were ahead of archeologists in Great Britain, who came on board around 1990. Clive Orton, one of the deans of quantitative archeology in Britain, argued that CA was the most important technique introduced into archeology in the 1980's [Orton, 1999, p. 32]. From Britain archeological CA migrated to the United States where it arrived shortly before to 2000. Duff [1996, p. 90] indicated, in an influential article from the mid 1990's, that CA was "not well established in Americanist literature". And, very recently, Smith and Neiman concurred: "CA has a long history of use by archeologists in continental Europe but its use by Americanist archeologists is both more recent and rare." [Smith and Neiman, 2007, p. 55].

There are several possible reasons why CA did not rapidly become popular in archeology in Britain and the United States. Most importantly, perhaps, archeological methodologists tend to look at statisticians for guidance, and in statistics CA was not really known until about 1980, despite the work of Hill [1974]. Except in France,

of course, but French statistics was relatively isolated from mainstream statistics. The dominant multivariate techniques applied in archeology were MDS and PCA (sometimes in the disguise of factor analysis). The most influential work in the area in the seventies was Hodson et al. [1971], which concentrated on the MDS techniques of Boneva, Kendall, and Kruskal. These are all forms of proximity analysis, but they differ from CA various ways.

LeBlanc [1975, p. 22] predicted, in a pioneering article: "Proximity analysis seems to hold a great deal of promise and will in all probability supplant all other seriation methods." If we interpret this prediction narrowly, in terms of the method that were available in 1975, it turned out to be incorrect, for reasons which are quite obvious in hindsight. Data, in archeology and elsewhere, come in many different forms. Sometimes we deal with cross tables, sometimes with incidence matrices, and sometimes with multivariate data that describe archeological objects in terms of a number of qualitative or quantitative variables. There is no reason to expect that a technique which is designed for one particular type of data will also work, or even be appropriate, for another type of data. A data analysis technique must obviously take the nature of the data into account, and forcing all data into a common "proximity"

format may not be an optimal strategy. But the basic advantages of proximity analysis mentioned by LeBlanc [1975, p. 22] are still right very much on target. "In the past, the basic goal of seriation has been to order a series of cultural units on the basis of an assumed single underlying variable, usually time. It is now possible to seriate units according to two or more variables by using a form of proximity analysis or MDS. This increases the power of seriation greatly, and among other advantages, it gives a much better idea of the fit of data to one variable (e.g. time alone) than have previous methods."

Because CA was rediscovered and reintroduced in different countries at different times, most archeological authors feel obliged to give some sort of introduction to the technique. These is even true for recent articles such as Poblome and Groenen [2003] and Smith and Neiman [2007]. Our discussion of CA differs in some respects from the ones traditionally encountered in archeology. In other respects it is quite standard. First, and this is actually quite common, we do not present the technique exclusively as a seriation method. There can be many different reasons why archeological sites are similar or dissimilar and, to quote Kruskal [1971], "Time is not the only dimension." Most CA plots are, of course, two-dimensional

maps in the plane, which already suggests more than one dimension may be relevant. Second, we discuss CA both as an exploratory technique and as a method of fitting a particular statistical model. And finally, we relate the least squares fitting of the CA model to the maximum likelihood fitting of the Exponential Distance (ED) model. Both ED and ordinary CA can be considered to be alternative, and closely related, forms of correspondence analysis.

1.2. **Types and Attributes.** LeBlanc [1975] compares *type seriation* and *attribute seriation*. See also Duff [1996]. We can discuss this comparison by distinguishing the different types of data that CA can be applied to. In a CA context attribute seriation corresponds to multiple correspondence analysis (MCA), treated in Gifi [1990, Chapter 3], and type seriation corresponds to simple CA, treated in Gifi [1990, Chapter 8]. Or, to translate this into software, attribute seriation corresponds with the `R` package `homals` [De Leeuw and Mair, 2008a], while type seriation corresponds with the package `anacor` [De Leeuw and Mair, 2008b].

LeBlanc [1975, p. 24] carefully distinguishes the terms “attribute”, “type”, “variable”, and “dimension”. Actuallly, he uses “variable” and “dimension” interchangeably, but is a probably a good idea to

reserve “dimension” for the axes in multidimensional representations in the data. A “variable” is then a formally defined aspect of the group of objects in the study. Each variable is measured in terms of a scale, and the mutually exclusive characteristics of the scale are called “attributes”. In the book by Gifi [1990], a variable is defined similarly as a mapping of the objects in the study into the categories of a variable. Defining a number of variables on a set of objects creates, in the terminology of the R software system [R Development Core Team, 2007], a “data frame”. More specific for archeology is the notion of a “type”, which Leblanc defines as “the existence of a non-random association between the attributes of two or more dimensions” [LeBlanc, 1975, p. 24]. Thus types are aggregations of attributes over different variables, and consequently they can be counted more easily, and are more susceptible to be treated with frequency-based techniques.

This discussion also makes it possible to compare CA with MDS and PCA. In MDS the first step is usually to derive some symmetric matrix of *similarities* between the sites, assemblages, proveniences, or cultural units. There are many ways to define similarities, and in

many cases the choice of a particular similarity measure is somewhat arbitrary. Moreover, instead of computing similarities between sites, we could also decide to compute similarities between the variables describing the artifacts found in the sites. A commonly used similarity measure between variables is the correlation coefficient. It is unclear how the MDS analysis of the sites and the MDS analysis of the variables are related. In PCA we usually start with a correlation matrix between variables, and then derive component loadings to describe the variables and component scores to describe the sites. This means PCA can be used to make a joint plot, a.k.a. a biplot [Gower and Hand, 1996]. Biplots are compelling ways to visualize multidimensional information, and as such they go beyond simple seriation.

One disadvantage of PCA that is often mentioned is that it assumes linear relations between the variables. This, however, is no longer true for modern non-linear versions of PCA, reviewed for example in De Leeuw [2006]. Moreover there is a close relationship between non-linear PCA and MCA, so close that in fact non-linear PCA can be carried out with the MCA package `homals` [De Leeuw and Mair, 2008a].

The correspondence analysis framework of Gifi [1990] gives one single class of techniques to analyse attribute matrices of artifacts by variables, frequency matrices of types by sites, and incidence matrices of types by sites. It is basically, to use a term from Benzécri's Analyse des Données, all a matter of "codage". One can code both types and sites as attributes of artifacts, and then the type by site frequency table is just the bivariate cross-table of those two variables.

One important advantage of CA and MCA over MDS and PCA is that they stay as close as possible to the original data, no matter if the data are frequencies or incidences or variables with attributes. There is no need to first choose a measure of similarity or correlation, and there is not need to aggregate data into correlation or product matrices. It is true that CA can be presented in terms of a particular measure of dissimilarity, the Benzécri distance. We will give such a presentation in this pape. But it is only one interpretation of the technique, and the Benzécri distances have a close connections with the familiar chi-squares that can be computed from the frequencies.

1.3. **Typical Archeological Applications.** We discuss some of the typical applications of CA in archeology in more detail, to illustrate where the technique may be appropriate and what archeologists look at.

In Bølviken et al. [1982] three data sets from the Stone Age in Northern Norway are used. The first one, from Iversfjord, uses thirty-seven lithic types in fourteen house site assemblages. Because of interpretational difficulties the analysis was repeated after grouping the thirty-seven types into nine tool categories. The joint plot in two dimensions of the house sites and tool categories is interpreted in terms of economic orientation and settlement permanence. The second example is for the Early Stone Age in the Varanger fjord area. The data counts frequencies of 16 functional tool types in 43 sites. Two-dimensional plots give a refinement interpreted in terms of earlier qualitative archeological hypotheses. The analysis was repeated grouping the tools into seven classes, yielding less informative results. In the third example CA was used to establish a chronology. Data came from a farm mount on the island of Helgøy in Troms. There are nineteen classes of artifacts

distributed over 15 excavation layers, carbon-dated from the fourteenth to the nineteenth centuries A.D. The analysis shows the layers mapped on a two-dimensional horseshoe curve. Projections on the curve can be used to reorder the rows and columns of the data matrix, producing a seriation closely corresponding with the one based on carbon-dating.

The article by Duff [1996] on micro-seriation compares attribute and type seriation, following LeBlanc [1975]. But whereas LeBlanc used multidimensional scaling for the type seriation, Duff used CA. Data are counts of six ceramic types in 40 proveniences in Pueblo de las Muertas, in the Zuni (Cibola) region of New Mexico, from the thirteenth to the fourteenth century A.D.. The two-dimensional CA solution exhibits a weak horseshoe, with lots of scatter around it, but produces essential the same ordering of the units as the MDS analysis of Leblanc.

An early application of CA to Americanist materials is Clouse [1999], who used CA to analyze artifacts found in excavations at the military settlement in Fort Snelling, Minnesota. Sites are eight defense buildings, eleven support buildings, and eight habitation buildings. At all sites artifacts are counted and classified in fourteen

groups, such as culinary, armament, commerce, furniture. Separate abundance matrices are given for defense, support, and habitation buildings and separate CA's are computed. Both joint plots, showing units and artifact groups in two dimensions, and unit-plots, which only show the units, are presented. Groupings of the units conform to what is expected on the basis of the Military Site Model, but provide more detailed information. Clouse [1999, p. 105] argues that CA makes expected and unusual features more clearly visible than the numerical summary given by the table.

The excellent paper by Smith and Neiman [2007] aims to compare frequency seriation, in the tradition of Ford [1952], with CA. They use two cases studies. In the first case study Gulf Coast area, near the Chattahoochee and Apalachicola Rivers, in Alabama, Georgia, and Florida. Data are from the Middle and Late Woodland periods (100 B.C. to A.D. 900). Ceramic data were collected at many sites, of which 29 were selected, because they had more than 80 painted sherds. The 29 sites were subdivided into 84 assemblages and the sherds were classified into 18 pottery types. Obviously it will be important for the eventual outcome of the technique how the artifacts and proveniences are grouped into rows and columns of the table. The CA of the 84 assemblages shows a very clear horseshoe

pattern, with a clear grouping of sites along the curve. "The CA results confirm what the clean seriation solution suggests: there is no significant source of variation in type frequencies other than time." [Smith and Neiman, 2007, p. 61] The analysis was repeated after removing some of the later assemblages. This smaller CA was validated (as a seriation method) by plotting CA scores against radiocarbon dates for selected sites.

The second case study in the Smith and Neiman article is from Kolomoki, a well-researched multimount site in southwestern Georgia. This is an intrasite analysis, not an analysis with multiple sites. The CA uses 20 assemblages and nine pottery types. Separate two-dimensional plots for assemblages and types shows no horseshoe, but a significant and interpretable second dimension. The CA solution shows effect, for example spatial ones, not detectable by the inherently one-dimensional frequency seriation. The first CA dimension is again validated as time, using radiocarbon data. We will use the same Kolomoki data set as one of our illustrative examples in this chapter.

2. Seriation

There is an interesting parallel historical development of what could broadly be called "seriation methods" in psychometrics, ecology, and archeology. The main steps in these development occur in the same order, but at different moments in time, not unlike archeological artifacts in different sites. Let's look at psychometrics first.

2.1. **Psychometrics.** In the 1940's, at the war department, Guttman [1944] discovered scalogram analysis, a method to simultaneously order attitude or achievement items (columns) and respondents (rows), with data in a binary data matrix. Initially scales were constructed by trial-and-error methods, in which row and columns of the binary data matrix were permuted to create the "consecutive ones" propety. More precisely, we look to order rows and columns in such a way that all ones are next to each other. This was done manually, using various ingineous devices. At the same time, the theory for principal components based computations was already available Guttman [1941, 1950]. In fact Guttman [1941] is the very first paper that rigorously defines MCA, and Guttman [1950] proves rigorously that the first MCA dimension provides the consecutive-one ordering for error-free data. The monumental

book by Coombs [1964] gave a systematic presentation of these heuristic pencil-and-paper techniques, applied to the various data matrices in proximity analysis. Although Coombs' conceptual framework is still relevant, the techniques were already superseded at the time of publication by computerized methods at the time the book appeared.

2.2. **Archeology.** Guttman's methods were published around 1950, almost simultaneously with Robinson [1951]. To discuss this work, we borrow some terminology from Kendall [1969]. An incidence matrix of, say, sites by types, is a *Petrie matrix* or *P-matrix* if in each column all ones occur consecutively. A non-negative symmetric matrix is a *Robinson matrix* or *R-matrix* if rows and columns are unimodal and attain their maximal values on the diagonal. By unimodal we mean that entries increase to a maximum and then decrease again. Similarities between sites whose incidence matrix is a P-matrix often form an R-matrix. Again, there is an interesting connection with psychometrics here. In the original definition of the Spearman model for general intelligence, dating back to 1904, a battery of tests satisfied the model if their correlation matrix was an R-matrix.

The notion of a P-matrix can be generalized to abundance matrices, i.e. to any matrix with no-negative entries. An abundance matrix is a *Q-matrix* if its columns are unimodal. That is the same as saying that the columns of the abundance matrix can be represented as a series of battleship plots, similar to the ones in Ford [1952] or Smith and Neiman [2007]. Many of the original archeological seriation techniques proposed by Petrie, Robinson, Ford, Hole and Shaw, and others take an incidence or abundance matrix and permute the sites in such a way that that it becomes a P-matrix or a Q-matrix. The permutation that is found then order the sites in time, i.e. it is a seriation. Ultimately, however, especially for large matrices finding optimal permutations is what is known in computer science as NP-hard, which basically means that the optimization problem, although finite, cannot be solved in a practical amount of time, even by the fastest computers Arlif [1995].

One way around the impractical computations involved with permutations is to use other related definitions of optimality. As we noted above, Guttman already proved in 1950 that CA can be used to find the optimal permutation to a P-matrix in the error-free case. For abundance matrices, see also Gifi [1990, Chapter 9], or Schriever [1983]. In fact, these papers prove more. They also

show that in the erro-free case the second dimension of the CA will be a quadratic function of the first, i.e. plotting the sites in the plane will show a quadratic curve.

Kendall [1971] and others later developed the well-know HORSHU program which applies MDS to smiliarties derived from abundance matrices, and then derives the order from the projection of the sites on the horseshoe or curvilinear arch. "We view the arch as a relatively benign indicator that the underlying data do, in fact, contain battleship-shaped curves." [Smith and Neiman, 2007, p. 60]

2.3. **Ecology.** In ecology the key concept is that of a "gradient". The emphasis in the data analysis is not on time, as in archeology, but on environmental characteristics. What is called "seriation" in archeology is called "ordination" in ecology [Gauch, Jr., 1982]. Plant or animal species do well under certain circumstances, and do best, for example, at some optimum level of wetness or altitude. Different species need different altitudes and/or different degrees of wetness. In ecology, of course, we have the major advantage that environmental gradients such as altitude can be directly measured. This is unlike psychometrics, where aptitude and attitude

are theoretical constructs, and unlike archeology, where direct information about the origin in time of an artifact is usually missing. So ecology has Direct Gradient Analysis, where we plot frequencies of species as a function of the gradient. In many cases we observe unimodal distributions, i.e. the abundance matrix is a Q-matrix.

Initially, same as in psychometrics and in archeology, ordination techniques used pencil-and-paper methods to reorder the rows and columns of the abundance matrix, or of derived similarity matrices with an Robinson structure [Whittaker, 1978]. This changed with the advent of the computer, and as in archeology and psychometrics, the ecologists turned to PCA and MDS for ordination, and to a host of measure of resemblance or similarity.

CA was introduced in ecology by Hill [1974] as “reciprocal averaging”. Ter Braak [1985] showed how CA was related to the unimodal response model, without going into precise mathematical detail. Ecologists initially were worried about the horseshoe, because they considered it an artifact, without any empirical significance. We now know more precisely where the arched structures come from, and we know that they indicate strong unidimensional effects. See

in particular Schriever [1985] or Van Rijckevorsel [1987]. We consequently tend to be pleased if we see a strong horseshoe, especially in archeology, where we have more reason perhaps to expect unidimensionality.

We will discuss the relationship between unimodal response models, in particular the Gaussian model of Ihm and van Groenewoud [1975], in more detail in Section 7 on the Exponential Distance Model.

3. Abundance Matrices

We now formalize some of the concepts we have mentioned in the introduction. Consider an $r \times c$ table N with *counts*. Rows correspond with r *sites*, columns with c *types*. Frequency n_{ij} indicates how often type j was found in site i. Such a matrix with counts N is called an *abundance matrix*. We also define the row sums $n_{i\bullet}$ and column sums $n_{\bullet j}$ of the table. The *grand total* $n_{\bullet\bullet}$ is the sum of all the counts in the table, which we will also abbreviate simply as n.

It should perhaps be mentioned that *presence-absence matrices* or *incidence matrices* are a special case of abundance matrices,

in which all entries of the table are either zero or one. An entry merely indicates if a type is present in a site or not. This means our discussion of abundance matrices also covers presence-absence matrices.

There is a more general type of data matrix, which is also quite common in archeology. Suppose the observation unit is an artifact such as a pottery sherd, a piece of obsidian, or maybe a fish bone. The units can be described in terms of a number of variables which can be either qualitative (categorical) or quantitative (numerical). The abundance matrix is a very special case of this, in which there are only two categorical variables used to describe the units, namely *site* and *type*.

The abundance data N can be coded as an $n \times 2$ matrix, where n is the grand total of the table, and where the first column is *site* and the second *type*. The table N is then the *cross-table*, or the *contingency table*, of the two variables. But clearly in a more general case variables such as size, color, weight, or composition could be used as well. For these more general multivariate data we need a technique such as MCA, also known as *homogeneity analysis*, [Gifi, 1990; Greenacre and Blasius, 2006]. Since the data analyzed in this

book are all of the simpler bivariate contingency table format, we shall not discuss MCA any further. As we mentioned in the introduction, it is the perfect technique for attribute-based seriation in the sense of LeBlanc [1975], in which we do not aggregate our data to types and assemblages, and to counts in a cross table.

3.1. **Examples.** Throughout the chapter we shall use two examples to illustrate the concepts of CA. The first example of an abundance matrix comes from a much larger matrix of sherd counts for sites by pottery types. All samples are from surface collections made ca. 1940 in Jalisco, Mexico by Kelley [1945].

This example is not a realistic application of CA because it is too small and too simple. The results of CA do not really add anything to what we can easily see by just looking at the table, but this very fact makes the example useful as an illustration of the basic concepts and calculations.

Insert Table 1 about here

The codes for the types, used as column headers, are

- AutPol: Autlan Polychrome;
- MiReBr: Miscellaneous Red on Brown, Buff;

- AuWhRe: Autlan White on Red;
- AltRed: Altillos Red Ware.

The sites are

- Site 21, Cofradia No. 1, and Site 34, Hacienda Nueva, are included in the Cofradia Complex (early);
- Site 23, Cofradia No. 3, and Site 37, Amilpa, are included in the Mylpa Complex (intermediate);
- Site 7, Mezquitlan, and Site 9, Altillos, are included in the Autlan Complex (late).

The second example are pottery data from the Kolomoki burial mounts in Georgia [Sears, 1956; Pluckhahn, 2003], analyzed previously with CA by Smith and Neiman [2007]. We already discussed these data in the introduction. There are 20 assemblages and 9 pottery types in the data.

4. Associated Matrices

With the abundance matrix we can associate several other matrices. In the first place there is the matrix P of *proportions*, whose elements are defined by

$$p_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}.$$

The matrix with proportions shows more clearly how the counts are distributed over the cells. Again the row marginals are $p_{i\bullet}$ and the column marginals are $p_{\bullet j}$.

Insert Table 2 about here

4.1. **Independence.** We say the row variable (site) and the column variable (type) are *independent* if $p_{ij} = p_{i\bullet}p_{\bullet j}$. Independence can be interpreted to mean that the body of the table does not give additional information, in fact all the information is contained in the marginals. If we know the relative frequencies of the sites and the types, then we can predict perfectly how many of each type there will be in each site.

We measure independence by what is called *inertia* in CA, borrowing a term from physics. Define the table Z of *Pearson residuals* with

$$z_{ij} = \frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}}}.$$

The elements of Z show the deviation between the observed proportion and the expected proportion on the hypothesis of independence (corrected for the standard error of the proportion). Positive elements indicate that we see more in the corresponding cell than we expect, negative elements mean that we see less. The *inertia* is

defined simply as

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} z_{ij}^2.$$

In the Kelley example the inertia is 0.9338, and the Pearson residuals are in Table 3.

Insert Table 3 about here

If the data are a random sample, and if types and sites are independent, then nX^2 is distributed as a chi-square random variable with $(r-1)(c-1) = 15$ degrees of freedom. In our example nX^2 is 1207.508. Moreover each of the $\sqrt{n}z_{ij}$ is approximately standard normal, i.e. it is what is commonly known as a z-score, and it can be tested for significance in the usual way. The z-scores are in Table 4.

Insert Table 4 about here

Clearly in the Kelley example the total inertia is far too big, the z-scores are mostly hugely significant, and the two variables *site* and *type* are very far from being independent. Of course in most archeological applications data very far from being a random sample, because we generally enumerate and classify all the artifacts

found in the site. Nevertheless we can still take inertia as a guideline to indicate how much structure there is in the data., or, more precisely, how much structure there is in the data that cannot be predicted from the marginals.

4.2. **Conditioning on Rows and Columns.** In archeological studies the hypothesis of independence is not the most natural way to look at abundance matrices. Independence is the appropriate concept if the contingency table results from a random sample from a discrete bivariate distribution, that is if we sample both sites and types. Usually, however, sites are not sampled. They are fixed either by design or by geographical circumstances.

What interests us really is to compare the distribution of types in the different sites that we have selected. Thus we are mainly interested in comparing the rows of the abundance matrix, because each row defines a distribution over types. Fortunately, the hypothesis of homogeneity of rows is mathematically equivalent to the hypothesis of independence. We can most easily see this by normalizing the rows, dividing each row by its row sum.

To keep our treatment symmetric, we also consider the case (less common in archeology) in which it may be interesting or appropriate to also compare the columns. Using the row and column sums, we can normalize the frequency table (or equivalently the table with proportions) by dividing the entries of the table by their row or column marginals. This defines two new tables, the first one conditioned by rows, the second conditioned by columns. The elements are defined by

$$p_{j|i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{p_{ij}}{p_{i\bullet}},$$
$$p_{i|j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{p_{ij}}{p_{\bullet j}}.$$

The hypothesis of independence $p_{ij} = p_{i\bullet}p_{\bullet j}$ can now be written in the two equivalent forms

$$p_{j|i} = p_{\bullet j},$$

$$p_{i|j} = p_{i\bullet},$$

which we can call *homogeneity of rows* and *homogeneity of columns*. Homogeneity of rows says that the probability distribution of types is the same for all sites. Homogeneity of columns says that the probability distribution of sites is the same for all types, which in

our context seems a less natural way of expressing the same basic mathematical fact.

Table 5 shows the distribution of types over each of the sites, with in the last row the distribution of types over all sites, i.e. the $p_{\bullet j}$. We have homogeneity if and only if all rows of the table, including the last row, are the same. Table 6 shows the distribution of sites over each of the types, with in the last column the distribution of sites over all types, i.e. the $p_{i\bullet}$. We have homogeneity if and only if all columns of the table, including the last column, are the same.

We can define appropriate measures of homogeneity of the rows and columns. These are again called *inertias* in CA. Thus there now is one inertia for each row, and one for each column. They are defined by

$$X^2_{i\bullet} = \sum_{j=1}^{c} \frac{(p_{j|i} - p_{\bullet j})^2}{p_{\bullet j}},$$

$$X^2_{\bullet j} = \sum_{i=1}^{r} \frac{(p_{i|j} - p_{i\bullet})^2}{p_{i\bullet}}.$$

Rows with a large inertia differ from the average row, i.e. the vector $p_{\bullet j}$ of column marginal proportions. And columns with a large inertia differ from the average column $p_{i\bullet}$.

Previously, we have defined the *total inertia*. Because of the simple relationship

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij} - p_{i\bullet}p_{\bullet j})^2}{p_{i\bullet}p_{\bullet j}} =$$

$$= \sum_{i=1}^{r} p_{i\bullet}X_{i\bullet}^2 = \sum_{j=1}^{c} p_{\bullet j}X_{\bullet j}^2$$

the total inertia is the weighted sum of the row and column inertias.

Insert Table 5 about here

Insert Table 6 about here

Under the hypothesis of random sampling from sites and homogeneity of rows, the $nX_{i\bullet}^2$ are distributed as chi-squares with $c - 1$ degrees of freedom. If we have random sampling and homogeneity of columns, the $nX_{\bullet j}^2$ are distributed as chi-squares with $r - 1$ degrees of freedom

5. Exploratory Correspondence Analysis

The basic purpose of exploratory CA is to make a *map of the types* and a *map of the sites*. By a "map" we mean a low-dimensional geometric representation. If we choose dimensionality equal to two, for instance, a map of the types consists of c points in the

plane, with one point corresponding with each type. If we choose dimensionality three, then a map of the sites consists of r points in three-dimensional space. Sometimes a one-dimensional map, which puts all sites on a straight line, is already enough to present the essential information in the table.

The location of the points in the map is not arbitrary, of course. If we make a two-dimensional map of the types, for example, we want to distances between the c points in the plane to be approximately equal to the distances between the c columns of the abundance matrix N. And similar for the map of the sites and the rows of N.

Distance on the map is defined in the usual way "as the crow flies". In other words, it is ordinary Euclidean distance. But distance between columns of the abundance matrix uses weights that takes the statistical stability of the cell counts into account. Specifically, in CA we use *Benzécri distances* (also known as *chi-square distances*). The squared Benzécri distance between row i and row k of table N is given by

$$\delta_{ik}^2 = \sum_{j=1}^{m} \frac{(p_{j|i} - p_{j|k})^2}{p_{\bullet j}},$$

and the squared Benzécri distance between column j and column ℓ of table N is

$$\delta^2_{j\ell} = \sum_{i=1}^{n} \frac{(p_{i|j} - p_{i|\ell})^2}{p_{i\bullet}}.$$

We give the squared Benzécri distances for the rows and columns in the Kelley example in Tables 7 and 8.

Insert Table 7 about here

Insert Table 8 about here

If we look more closely at Table 7 we can already predict what CA will do. If we want a geometric representation in which the distances approximate the Benzécri distaces, then it is pretty clear how such a representation would look. The Benzécri distances between sites 21 and 34 and between sites 23 and 37 are almost zero. Thus in a map sites 21 and 34 will coincide, and sites 23 and 37 will also coincide. Sites 9 and 7 are close as well, and $(21, 34)$ is about equally distant from the two groups $(7, 9)$ and $(23, 37)$. A two-dimensional map will thus look like an isosceles triangle with the three groups of sites at the edges. The shorter side is somewhere around $\sqrt{2}$ or $\sqrt{3}$, the two longer sides are around $\sqrt{6}$. We also see that it will in general be impossible to map the distance information on a straight line, because in that case we would have to

let $(7, 9)$ coincide with $(23, 37)$. In this small example we can easily see what a map would look like, but in a larger examples, such as the Kolomoki one, this becomes much more complicated. That is why we have CA, which approximates the Benzécri distances by the Euclidean distances in the map in a precise way.

In CA we approximate Benzécri distances *from below*. Let us explain this concept. In any CA map of the sites, for instance, we will always have $d_{ik} \leq \delta_{ij}$, where d_{ik} is Euclidean distance between points i and k on the map. More precisely, CA constructs a sequence of maps, the first one has only one dimension, the second one has two, and so on. The final map has $t = \min(r - 1, c - 1)$ dimensions, i.e. 3 in the Kelley example and 8 in the Kolomoki example. The maps are *nested*, in the sense that the projections on the first dimension of all the maps is identical to the one-dimensional map, and the projection on the plane of the first two dimensions for all maps with dimension at least two is equal to the two-dimensional map. And so on. If $d_{ik}^{(s)}$ are the distances is the s-dimensional map, with $1 \leq s \leq t$, then

$$d_{ik}^{(1)} \leq d_{ik}^{(2)} \leq \cdots \leq d_{ik}^{(t)} = \delta_{ik}.$$

Thus the t-dimensional map has distances exactly equal to the Benzécri distances. Maps in fewer dimensions approximate the distances, and the approximation becomes better, for each of the distances, when the dimensionality increases. Approximation is from below, because map distances are always smaller than Benzécri distances, no matter what the dimensionality of the map is. Of course the same reasoning applies to Benzécri distances between columns and the CA map for types.

The map does not only approximate Benzécri distances between sites or types, it also approximates the inertias of the sites and the types. In the sites map, for instance, the inertia is approximated (from below, as usual) by the distance of the site to the origin of the map. Or, equivalently, by the length of the vector corresponding with the site. This means that a site that differs little from the average site, and thus has a small inertia, will be close to the origin of the map. And sites that are different from the others will be tend to be in the periphery of the map. As a consequence it can happen quite easily that the center of the map, the area near the origin, is somewhat cluttered with sites that are similar to the average site.

A CA program (we use De Leeuw and Mair [2008b]) typically takes the abundance matrix and the desired dimensionality of the map as its arguments. It then outputs coordinates for the maps of the row objects (sites) and the column objects (types). In addition it can provide a variety of plots, and it provides a *decomposition of the inertia*. This type of decomposition is familiar from PCA. Consider the weighted squared length of the projections of the site points on the first dimension, on the second dimension, and so on. This decomposes the total inertia of the vectors into a component die to the first dimension, to the second dimension, and so on. By dividing the components by the total, we can say that a certain percentage of the inertia is "explained" by the first dimension, another, smaller, percentage by the second dimension, and so on. Ultimately there are $t = \min(r-1, c-1)$ dimensions, and each of them takes care of a certain decreasing percentage of the total inertia.

CA can also make *joint maps*, or *biplots*, in which we basically take the site plot and the type plot and put them on top of each other. We then have a plot in which types will tend to be close to sites in which they occur more frequently than one would expect on the basis of the marginals. We say "tend to", because there

is not Benzécri distance defined between a site and a type, and thus there is no approximation in some well-defined mathematical sense. The CA program basically lets the user make four choices for the joint plot. The first one is to put the two Benzécri plots on top of each other. Distances between sites, and distances between types, approximate Benzécri distances, but distances between sites and types have no simple relation to the data. The second option, which is called Goodman scaling in the program, is to adjust the length of the site and type vectors in such a way that their inner product approximates the Pearson residual. Unfortunately this invalidates the interpretation of site and type distances as approximations of Benzécri distances. The last two options use the *centroid principle*. We can take the Benzécri map for the sites, and then plot the types by taking weighted averages (centroids) of the sites, using the frequencies of the types in those sites as weights. This produces a joint plot in which site distances approximate Benzécri distances. The locations of the types in the plot again only differ in vector length from the locations in the Benzécri type plot. Type distances cannot be interpreted as approximating Benzécri distances between types any more, but they do have a clear geometric interpretation as weighted averages of site points. By symmetry there

is a second centroid principle, in which we use the Benzécri type plot and then plot the sites as weighted averages of types.

5.1. **Kelley.** Let us illustrate exploratory CA with the small Kelley example. The two-dimensional maps for sites and types from CA are in Figure 1.

Insert Figure 1 about here

As expected, in the sites map, we see the three clusters of points at the vertices of a triangle. As we know, the one-dimensional map is simply the projection of all points on the horizontal axis.

Insert Figure 2 about here

In Figure 2(a) we see the approximation of the Benzécri distances between sites in one dimension, and in Figure 2(b) in two dimensions. Benzécri distances are on the horizontal axis, Euclidean map distances on the vertical axis. Approximation from below means that all points are below the 45 degree line of perfect fit. But, as we can see, fit in two dimensions is already almost perfect. In one dimension some of the larger Benzécri distances, in particular those between $(21, 34)$ and $(23, 37)$ are seriously underestimated.

We finally show the chi-square decomposition for the Kelley example. Not surprisingly, the two first dimensions account for 97% of the total inertia, and the third dimension is of very little importance.

Insert Table 9 about here

5.2. **Kolomoki.** We now apply CA to the Kolomoki data,our more realistic example. The chi-square decomposition is given Table 10. Two dimensions account of 80% of the inertia, three dimensions for almost 90%. The CA maps for the types in two and three dimensions are given in Figure 3 and Figure 4. Again, the two-dimensional map is just the projection of the three-dimensional map on the horizontal plane (except for a possible rotation). Note that the points in the two-dimensional maps are center of ellipses of varying sizes. These ellipses are 95% confidence regions for the points. Confidence region computations, which are done in De Leeuw and Mair [2008b], are based on the assumption that the abundnaces are a large random sample from a population. As with chi-square, this assumption may not be appropriate in archeological examples, but, also as with chi-square, the size of the ellipses

does give a useful representation of variability. We see larger ellipses for outlying points, which generally correspond with smaller abundances, and we see examples of overlapping ellipses for sites or type that cannot really be distinguished.

Insert Figure 3 about here

For the interpretation of the two-dimensional Kolomoki results, we refer to the experts Smith and Neiman [2007]. The third dimension does not add much (only 9% of the total inertia), but it does allow us to better approximate some of the larger Benzécri distances. In particular the third dimensions emphasizes the differences between the outliers T9 and (T1,T18).

Insert Figure 4 about here

If we continue to add dimensions, we will probably see each new dimension take care of a group of the large Benzécri distances, which are still seriously underestimated in three dimensions.

Insert Figure 5 about here

Insert Table 10 about here

6. Frequently Asked Questions

There are various variations of CA that naturally come to mind. We have not applied them in our example, but we briefly mention them for completeness. On can wonder, for example, if approximation from below is such a good idea. It seems ovious that better approximation of the Benzécri distances is possible if we allow some of the map distances to overestimate, and other to underestimate. This idea is exploited in [De Leeuw and Meulman, 1986]. The idea, basically, it to compute Benzécri distances first, and then apply multidimensional scaling to these distances.

A second question is if there are suitable alternatives to the Benzécri distances. Remember that Benzécri distances are used because we correct the proportions for their standard errors, on the assumption of independence. Benzécri distances have a natural connection to chi-square, to weighted sum of squares, and thus to Euclidean distance. Alternative method to weight the proportions are indeed possible, as in the spherical CA of Domingues and Volle [1980], but generally the connection with Euclidean geometry becomes less transparant.

And finally, we can get away from the interpretation of abundance matrices in terms of relative frequencies. Instead we can think of them as *compositional data*. Each row is a vector of proportions, adding up to one, but the proportions may come from a chemical analysis of samples, and may not come from counts. Compositional data are very common in chemometrics and the earth sciences, and also quite common in archeology. Variations of principal component analysis for compositional data similar to, but not identical with, CA are discussed in the monograph by Aitchison [2003].

7. Exponential Distance Models

In ecology [Ihm and van Groenewoud, 1975; Ter Braak, 1985], and to some extent in archeology, much attention has been paid to the Gaussian Ordination Model (GOM). The model says that for site i and species j the expected value of the abundance is

$$\mathbf{E}(f_{ij}) = \alpha_i \beta_j \exp(-\frac{1}{2}\left(\frac{x_i - y_j}{s_j}\right)^2).$$

Thus sites and types can be scaled on a common one-dimensional scale. Abundance f_{ij} is, except for the marginal row and column effects α_i and β_j, related to the distance between the scale-value

of site i and the scale-value of type j. More precisely, a type will be abundant in sites whose scale value is close to the type's scale value, and it will be largest if type and site coincide on the scale. Rows of the abundance matrix will be unimodal: they have a single peak and then level off in both directions. Or, using Kendall's terminology, they are Q-matrices. Again, except for the marginal effects, the same thing is true for the columns. Thus, if the model fits, we can reorder the sites and types in such a way that both rows and columns of the abundance matrix are unimodal.

Fitting. ML. RC model.

The GOM can be generalized easily to more than one dimension.

$$\mathbf{E}(f_{ij}) = \alpha_i \beta_j \exp(-\frac{1}{2} \sum_{s=1}^{p} (x_{is} - y_{js})^2).$$

This model is unimodal in a more general geometrical sense. The response curves in the plane, if $p = 2$, have a single peak and level off in all directions. There are many ways in which the GOM can be fitted to abundance matrices. Nor surprisingly, there have been contributions from both psychometrics and ecology. For a recently proposed technique, and a good overview of earlier work, we refer to De Rooij and Heiser [2005].

We can simplify this, by expanding the square and collecting terms, to the more simple but equivalent form

$$\mathbf{E}(f_{ij}) = \tilde{\alpha}_i\tilde{\beta}_j \exp(\sum_{s=1}^{p} x_{is}y_{js}).$$

This shows we expand the abundances into marginal effects and an interaction terms which is an inner product of row and column effects. This is actually quite close to CA. For small arguments we have $\exp(x) \approx 1 + x$, and consequently

$$\mathbf{E}(f_{ij}) \approx \tilde{\alpha}_i\tilde{\beta}_j \left(1 + \sum_{s=1}^{p} x_{is}y_{js}\right).$$

8. Discussion

This chapter could be called "the many faces of Correspondence Analysis". It tries to provide various interpretational frameworks to look at CA plots, in terms of distances, centroids, association models, and chi-square. It also shows how the same models and techniques appear in many different disciplines, often under different names, and that combining ideas from these disciplines gives additional possibilities of interpretation.

We have also discussed the EDM model, in its various disguises as the GOM or the RC-model. It can be used to embed a form of CA

into a maximum likelihood framework and to shift the emphasis from multivariate exploration to model testing.

Archeologists not familiar with CA can use this chapter to look at previous examples in their discipline, and to think in a different way about abundance and incidence matrices. We have tried to emphasize the continuity between CA and previous seriation methods used in archeology.

As we have indicated, there are convenient free R packages available for CA. We mentioned `hoaml s` and `anacor`, but in De Leeuw and Mair [2008b] other available packages are discussed as well. All standard statistical systems, such as SAS, SPSS, and Stata, also have CA methods as either built-ins or add-ons.

References

J. Aitchison. *The Statistical Analysis of Compositional Data*. The Blackburn Press, Caldwell, New Jersey, 2003.

B.V. Arlif. The Archeological Seriation Problem. Master's thesis, Instutute of Computer Science, University of Copenhagen, Denmark, 1995.

M.J. Baxter. *Exploratory Multivariate Analysis in Archeology*. Edinburgh University Press, Edinburgh, 1994.

E. Beh. Simple Correspondence Analysis: A Bibliographic Review. *International Statistical Review*, 72:257–284, 2004.

J.P. Benzécri. *Analyse des Données: Taxonomie*, volume 1. Dunod, Paris, 1973a.

J.P. Benzécri. *Analyse des Données: Correspondances*, volume 2. Dunod, Paris, 1973b.

E. Bølviken, E. Helskog, K. Helskog, I.M. Holm-Olsen, L. Solheim, and R. Bertelsen. Correspondence Analysis: an Alternative to Principal Components. *World Archeology*, 14:41–60, 1982.

C.C. Clogg and E.S. Shihadeh. *Statistical Models for Ordinal Variables*. Number 4 in Advanced Quantitative Techniques in the Social Sciences. Sage Publications, Thousand Oaks, CA, 1994.

R.A. Clouse. Interpreting Archeological Data Through Correspondence Analysis. *Historical Archeology*, 33:90–107, 1999.

C. H. Coombs. *A Theory of Data*. Wiley, 1964.

J. De Leeuw. Nonlinear Principal Component Analysis and Related Techniques. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, 2006.

J. De Leeuw. On the Prehistory of Correspondence Analysis. *Statistica Neerlandica*, 37:161–164, 1983.

J. De Leeuw and P. Mair. Homogeneity Analysis in R: The package homals. *Journal of Statistical Software*, 2008a.

J. De Leeuw and P. Mair. Simple and Canonical Correspondence Analysis Using the R Package anacor. *Journal of Statistical Software*, 2008b.

J. De Leeuw and J.J. Meulman. Principal Component Analysis and Restricted Multidimensional Scaling. In W. Gaul and M. Schader, editors, *Classification as a Tool of Research*, pages 83–96, Amsterdam, London, New York, Tokyo, 1986. North-Holland.

M De Rooij and W.J. Heiser. Graphical Representations and Odds Ratios in a Distance Association Model for the Analysis of Cross-Classified Data. *Psychometrika*, 70:99–122, 2005.

D. Domingues and M. Volle. L'Analyse Factorielle Sphérique. In E. Diday, L. Lebart, J. Pagès, and R. Tomassone, editors, *Data Analysis and Informatics*, volume I, Amsterdam, Netherlands, 1980. North Holland Publishing Company.

A.I. Duff. Ceramic Micro-Seriation: Types or Attributes. *American Antiquity*, 61:89–101, 1996.

J.A. Ford. Measurements of Some Prehistoric Design Developments in the Southeastern States. *Anthropological Papers of the American Museum of Natural History*, 44(3):313–384, 1952.

H.G. Gauch, Jr. *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge, U.K., 1982.

A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, England, 1990.

Z. Gilula and S.J. Haberman. Canonical Analysis of Contingency Tables by Maximum Likelihood. *Journal of the American Statistical Association*, 81:780–788, 1986.

L.A. Goodman. Simple Models for the Analysis of Association in Cross-classifications Having Ordered Categories. *Journal of American Statistical Association*, 74:537–552, 1979.

J.C. Gower and D.J. Hand. *Biplots*. Number 54 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1996.

M. Greenacre and J. Blasius, editors. *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall, 2006.

L. Guttman. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, editor, *The Prediction of Personal Adjustment*, pages 321–348. Social Science Research Council, New York, 1941.

L. Guttman. A Basis for Scaling Qualitative Data. *American Sociological Review*, 9:139–150, 1944.

L. Guttman. The Principal Components of Scale Analysis. In S.A. Stouffer and Others, editors, *Measurement and Prediction*. Princeton University Press, Princeton, 1950.

M.O. Hill. Correspondence Analysis: a Neglected Multivariate Method. *Applied Statistics*, 23:340–354, 1974.

F.R. Hodson, D.G. Kendall, and P. Tăutu, editors. *Mathematics in the Archeological and Historical Sciences*, Edinburgh, 1971. Edinburgh University Press.

P. Ihm and H. van Groenewoud. A Multivariate Ordering of Vegetation Data Based on Gaussian Type Gradient Response Curves. *The Journal of Ecology*, 63:767–777, 1975.

I. Kelley. *The Archeology of the Autlán-Tuxcacuesco Area of Jalisco. I: The Autlán Zone*, volume 26 of *Ibero-Americana*. University of California Press, 1945.

D.G. Kendall. Incidence Matrices, Interval Graphs, and Seriation in Archeology. *Pacific Journal of Mathematics*, 28:565–570, 1969.

D.G. Kendall. Abundance Matrices and Seriation in Archeology. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 17:104–112, 1971.

J.B. Kruskal. Multi-dimensional Scaling in Archeology: Time is not the Only Dimension. In F.R. Hodson, D.G. Kendall, and P. Tăutu,

editors, *Mathematics in the Archeological and Historical Sciences*, pages 119–132, Edinburgh, 1971. Edinburgh University Press.

S.A. LeBlanc. Micro-Seriation: A Method for Fine Chronologic Differentiation. *American Antiquity*, 40:22–38, 1975.

J. Müller and A. Zimmerman, editors. *Archeology and Correspondence Analysis. Examples, Questions, Perspectives.*, volume IA 23 of *Internationale Archäologie*. Verlag Marie Leidorf, Rahden, Germany, 1997.

C. Orton. Plus ça Change ? 25 Years of Statistics in Archeology. In L. Dingwall, S. Exon, V. Gaffney, S. Laflan, and M. van Leusen, editors, *Archeology in the Age of the Internet*, Oxford, 1999. Archopress.

T.J. Pluckhahn. *Kolomoki: Settlement, Ceremony, and Status in the Deep South, A.D. 350-750.* University of Alabama Press, Tuscaloosa, Alabama, 2003.

J. Poblome and P.J.F. Groenen. Constrained Correspondence Analysis for Seriation of Sagalassos Tablewares. In M. Doerr and A. Sarris, editors, *Computer Applications and Quantitative Methods in Archaeology*, pages 301–306. Hellinic Ministry of Culture, 2003.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing,

Vienna, Austria, 2007. URL `http://www.R-project.org`.

W.S. Robinson. A Method for Chronologically Ordering Archeological Deposits. *American Antiquity*, 16:293–301, 1951.

B.F. Schriever. *Order Dependence*. PhD thesis, University of Amsterdam, The Netherlands, 1985. Also published in 1985 by CWI, Amsterdam, The Netherlands.

B.F. Schriever. Scaling of Order-dependent Categorical Variables with Correspondence Analysis. *International Statistical Review*, 51:225–238, 1983.

W.H. Sears. *Excavations at Kolomoki: Final Report*. University of Georgia Press, Athens, Georgia, 1956.

K.Y. Smith and F.D. Neiman. Frequency Seriation, Correspondence Analysis, and Woodland Periodic Ceramic Assemblage Variation in the Deep South. *Southeastern Archeology*, 26(1):47–72, 2007.

C.J.F. Ter Braak. Correspondence Analysis of Incidence and Abundance Data: Porperties in Terms of a Unimodal Response Model. *Biometrics*, 41:859–873, 1985.

J.L.A. Van Rijckevorsel. *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. PhD thesis, University of Leiden, The Netherlands, 1987. Also published in 1987 by DSWO-Press, Leiden, The Netherlands.

R.H. Whittaker, editor. *Ordination of Plant Communities*. Dr. W. Junk BV, The Hague, Netherlands, 1978.

T.W. Yee. A New Technique for Maximum Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, 74:685–701, 2004.

TABLE 1. Abundance Matrix from Kelley

	Type				
	AutPol	MiReBr	AuWhRe	AltRed	
21	8	14	0	0	22
34	19	35	0	0	54
23	138	6	0	1	145
37	299	11	0	2	312
9	102	12	22	271	407
7	34	14	59	246	520
	600	92	81	520	1293

TABLE 2. Proportions Matrix Kelley

	Type				
	AutPol	MiReBr	AuWhRe	AltRed	
21	0.006	0.011	0.000	0.000	0.017
34	0.015	0.027	0.000	0.000	0.041
23	0.107	0.005	0.000	0.001	0.112
37	0.231	0.009	0.000	0.002	0.241
9	0.079	0.009	0.017	0.210	0.315
7	0.026	0.011	0.046	0.190	0.273
	0.464	0.071	0.063	0.402	1.000

TABLE 3. Pearson Residuals Kelley

	Type			
	AutPol	MiReBr	AuWhRe	AltRed
21	-0.02	+0.28	-0.03	-0.08
34	-0.03	+0.44	-0.05	-0.13
23	+0.24	-0.04	-0.08	-0.21
37	+0.35	-0.07	-0.12	-0.31
9	-0.18	-0.09	-0.02	+0.23
7	-0.28	-0.06	+0.22	+0.24

TABLE 4. z-scores Kelley

	Type			
	AutPol	MiReBr	AuWhRe	AltRed
21	-0.69	+9.94	-1.17	-2.97
34	-1.21	+15.90	-1.83	-4.66
23	+8.62	-1.34	-3.01	-7.50
37	+12.81	-2.38	-4.42	-11.02
9	-6.32	-3.15	-0.69	+8.39
7	-10.14	-2.22	+7.84	+8.73

TABLE 5. Conditioning on the rows in Kelley

	Type				
site	AutPol	MiReBr	AuWhRe	AltRed	$X^2_{i\bullet}$
21	0.36	0.64	0.00	0.00	4.98
34	0.35	0.65	0.00	0.00	0.04
23	0.95	0.04	0.00	0.01	0.11
37	0.96	0.03	0.00	0.01	0.24
9	0.25	0.03	0.05	0.52	0.31
7	0.10	0.04	0.17	0.47	0.27
$p_{\bullet j}$	0.46	0.07	0.06	0.40	0.93

TABLE 6. Conditioning on the columns in Kelley

	Type				
site	AutPol	MiReBr	AuWhRe	AltRed	$p_{i\bullet}$
21	0.01	0.15	0.00	0.00	0.02
34	0.03	0.38	0.00	0.00	0.04
23	0.23	0.07	0.00	0.00	0.11
37	0.50	0.12	0.00	0.00	0.24
9	0.17	0.13	0.27	0.52	0.31
7	0.06	0.15	0.73	0.47	0.27
$X^2_{\bullet j}$	0.64	4.06	1.18	0.68	0.93

TABLE 7. Squared Benzécri Distances Rows (Sites)

	21	34	23	37	9	7
21	0.000					
34	0.002	0.000				
23	5.721	5.950	0.000			
37	5.841	6.072	0.001	0.000		
9	6.353	6.550	2.188	2.208	0.000	
7	6.812	6.999	3.207	3.233	0.259	0.000

TABLE 8. Squared Benzécri Distances Columns (Types)

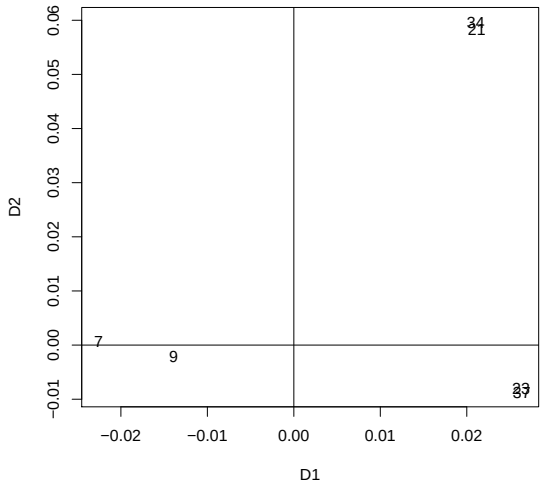
	AutPol	MiReBr	AuWhRe	AltRed
AutPol	0.000			
MiReBr	4.921	0.000		
AuWhRe	3.221	6.203	0.000	
AltRed	2.539	5.780	0.436	0.000

TABLE 9. Chi-square Decomposition Kelley

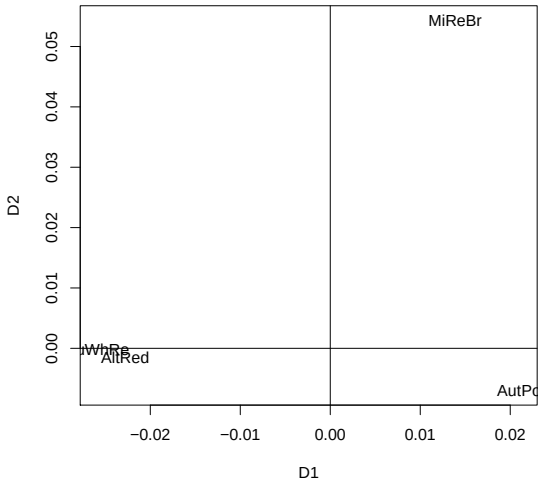
	X^2	%	Cum %
1	787.9	0.65	0.65
2	390.0	0.32	0.97
3	29.6	0.03	1.00
Total	1207.5		

TABLE 10. Chi-square Decomposition Kolomoki

	X^2	%	Cum %
1	1018.8	0.63	0.63
2	261.6	0.16	0.79
3	144.7	0.09	0.88
4	128.0	0.08	0.96
5	38.6	0.02	0.98
6	17.9	0.01	0.99
7	9.0	0.01	1.00
8	3.8	0.00	1.00
Total	1622.5		

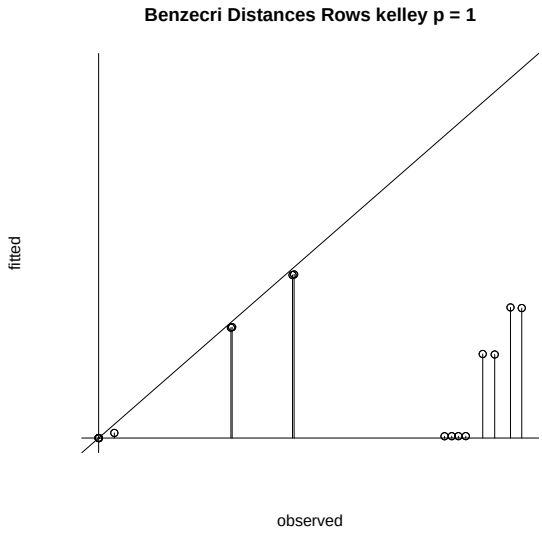


(a) Sites

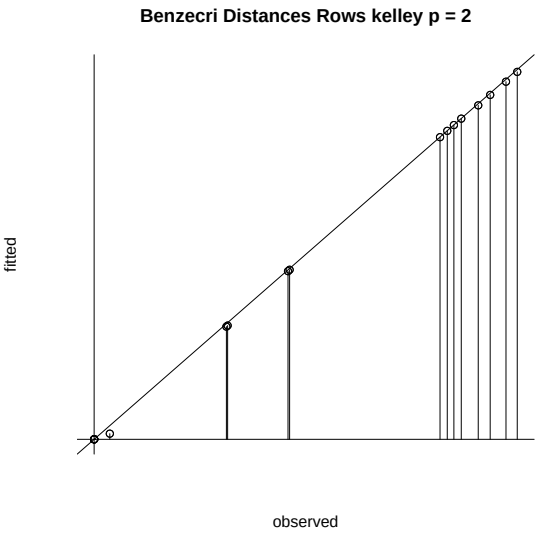


(b) Types

FIGURE 1. Two-dimensional CA Map for Kelley

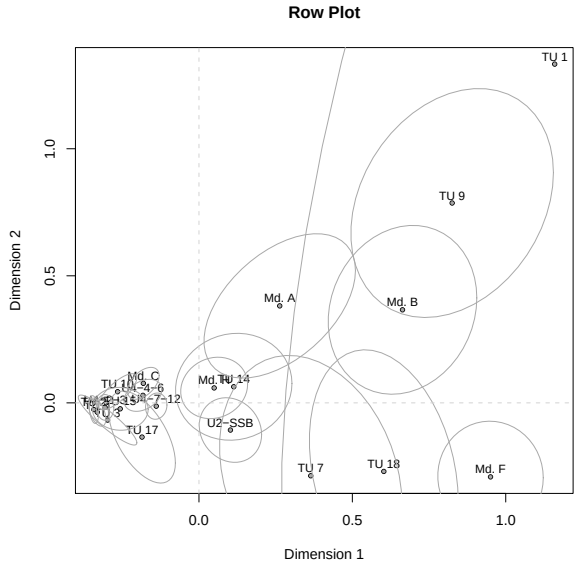


(a) One Dimension

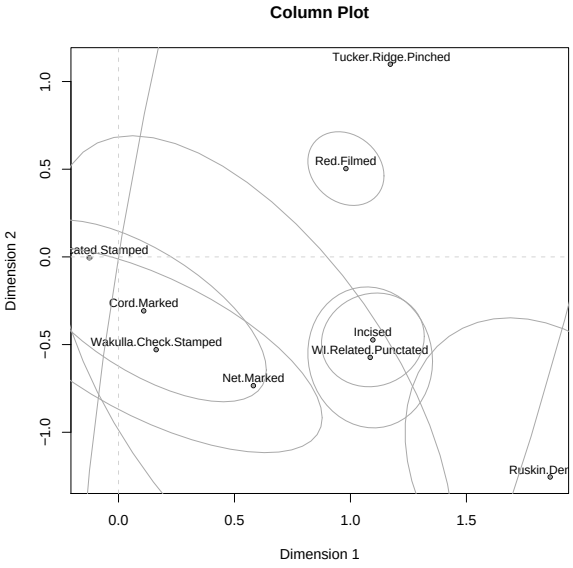


(b) Two Dimensions

FIGURE 2. Approximation of Benzécri Distances for Kelley



(a) Rows



(b) Columns

FIGURE 3. CA Maps of Kolomoki

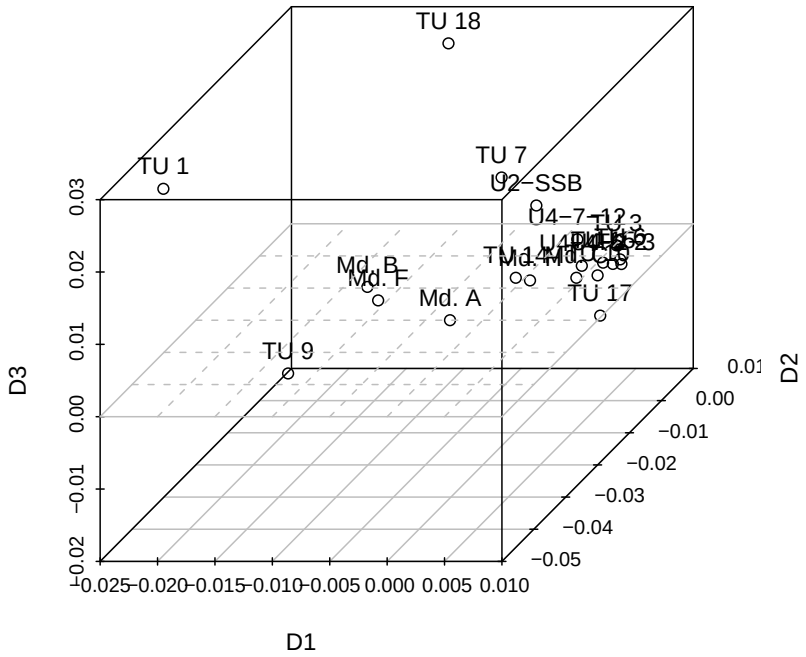
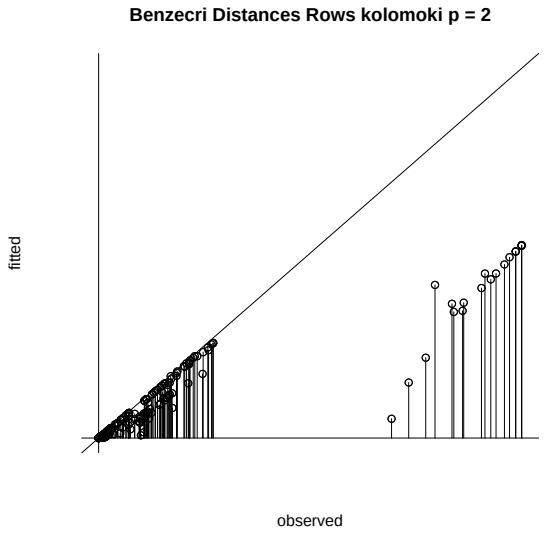
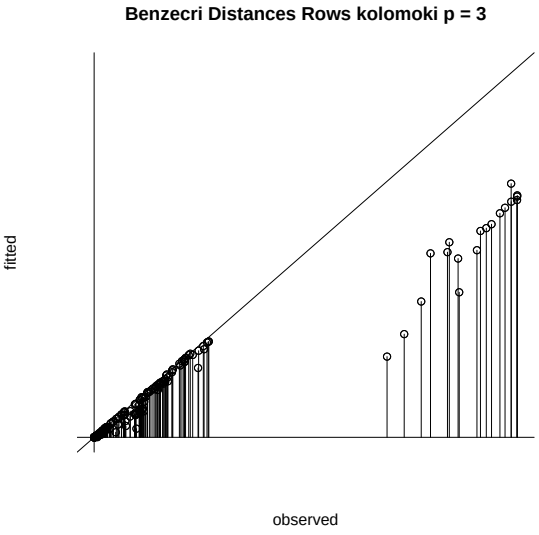


FIGURE 4. Three Dimensional Map of Kolomoki

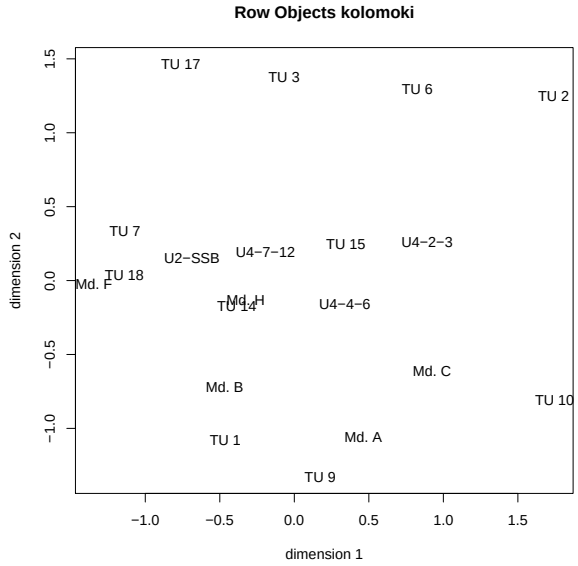


(a) Two Dimension

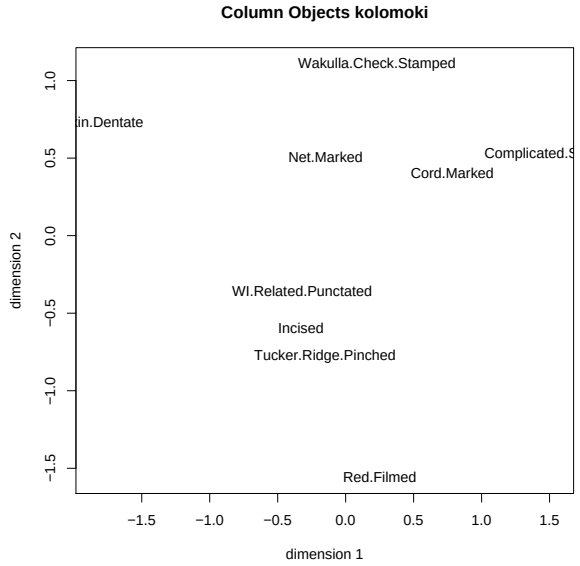


(b) Three Dimensions

FIGURE 5. Approximation of Benzécri Distances for Kolomoki



(a) Rows



(b) Columns

FIGURE 6. EDM Maps of Kolomoki