Archeologia e Calcolatori
24, 2013, 25-53

AN R SCRIPT TO FACILITATE CORRESPONDENCE ANALYSIS. A GUIDE TO THE USE AND THE INTERPRETATION OF RESULTS FROM AN ARCHAEOLOGICAL PERSPECTIVE

1. Introduction

In a recent issue of this journal, M.J. Baxter and H.E.M. Cool published an interesting article on the use of Correspondence Analysis (hereafter, CA) in archaeology (Baxter, Cool 2010). That article, as well as previous ones by the same authors (e.g., Cool, Baxter 1999), has clearly underscored the usefulness of the technique for the interpretation of the complex datasets archaeologists often happen to deal with. The graphical display of rows and columns of a contingency table enables the analyst to search the likely dimension of the data and explore different trends of variability, allowing hidden patterns to emerge.

The use of CA has steadily increased in social science (see, e.g., the papers in Blasius, Greenacre 1998) as well as in archaeology. Even though in the latter field CA has been slow in gaining popularity, with the exception of early groundbreaking studies (Bølviken *et al.* 1982; Djindjian 1985; Madsen 1989; Gillis 1990), today it is used for many purposes, including intrasite activity areas research (Kuijt, Goodale 2009; Alberti 2012, 2013), burial assemblages analysis (Wallin 2010), on-site distribution of faunal remains (Potter 2000; Morris 2008), distribution of drinking pottery types in the context of cultures contact (Pitts 2005), stratigraphy and formation processes (Mameli *et al.* 2002; Pavùk 2010), seriation and chronology (Bellanger *et al.* 2008; Peeples, Schachner 2012).

Given the relevance and utility of CA, Baxter and Cool are right in advocating a more widespread use of the technique (Baxter, Cool 2010, 213, 225), and their effort in providing a detailed guide to perform CA in the free statistical R environment (Ihaka, Gentleman 1996) is very welcome. In fact, all the main commercial statistical programs can perform CA, but their price is generally far beyond the budget of the average user, let alone students willing to approach the technique to analyse data on their own. On the contrary, many packages that perform CA are freely available in R, each with different features as far as graphical output and analytical tools are concerned: see, e.g., the "anacor" (de Leeuw, Mair 2009), "ca" (Nenadic, Greenacre 2007), and "FactoMineR" packages (Lê *et al.* 2008; Husson *et al.* 2011), or those (namely, "MASS", "ade4", and "vegan") used by Baxter and Cool (2010, references therein). The availability of many different tools offers to the user the possibility to choose the one(s) he considers more appropriate for his analytical tasks.

As for the choice I made, the decision to focus on the "ca" and "FactoMineR" packages rests on both matter of personal taste and on the fact that extensive literature does exist allowing the users to go deeper into the details of both packages (Greenacre 2007, 213-258; Nenadic, Greenacre 2007; Lê *et al.* 2008; Husson *et al.* 2011). Remarkably, video tutorials on the use of "FactoMineR" have been made available by F. Husson himself and can be easily found on his YouTube channel (http://www.youtube.com/user/HussonFrancois).

2. Aim of the article

The aim of the article is twofold, namely:

– to expand Baxter and Cool's sensible idea of benefitting from the flexibility of the R environment to perform CA;
– to make a step forward in the direction of freeing the user from manually entering long pieces of R code; in this respect, an R script will be proposed. It will be soon made available both on-line (http://uniud.academia.edu/GianmarcoAlberti/) and from the author upon request. A video tutorial on YouTube is also planned.

It is not the intention of this article to instruct the readers on the coding needed to perform CA. As a matter of fact, I do not want to replicate what already exists in literature: many scholars have already focused on line-by-line tutorials of CA in R (Nenadic, Greenacre 2007; Greenacre 2007, 213-258; Baxter, Cool 2010; Husson *et al.* 2011, 59-126; Glynn in press). Rather, this work is intended for archaeologists with no or scant knowledge of R yet willing to use it to perform CA. For this reason, the article concentrates on the analysis' output rather than on the way to obtain it from R. More experienced R users will already be familiar enough to grasp the script on their own and to use (or even modify) it according to their personal taste and specific needs.

What are the advantages of the script? It allows the user to:

– pick the best (or, at least, what I consider as such) from the aforementioned two R packages developed by leading scholars in CA computation, in order to provide a set of CA statistics and graphical outputs relevant to the analysis of data;
– provide a textual summary of the CA output statistics;
– provide graphs (some of them not native to the packages) that are important for CA interpretation;
– provide the possibility to compare four different criteria for the selection of an optimal dimensionality of the CA solution; in this respect, the Malinvaud's test (full discussion and references in § 4.2.2) has been implemented for the first time in R, at the best of my knowledge.

The use of CA for seriation purposes is not the main concern here, and on this topic the reader is referred to the available literature (Weller, Romney 1990, 76-83; Baxter 1994, 118-123; Kjeld Jensen, Høilund Nielsen 1997; Smith, Neiman 2007; Baxter, Cool 2010, 218-220; Peeples, Schachner 2012).

In what follows, first, a brief introduction to CA will be sketched up. A full account of its theoretical and computational underpinnings can be found in Greenacre (2007) and, from an archaeological perspective, in Baxter (1994, 100-139) and Shennan (1997, 308-341). Later, the discussion of a detailed (fictional) worked example will bring us into the core of the article's argument, introducing the advantages of the author's R script, i.e. a time-saving sequence of commands that can be executed by any user at any time. It allows the execution of CA in R without the need to manually enter long pieces of code, allowing the user to concentrate on the analysis' results rather than, as said, on the ways to obtain them from R. Whereas the fictional worked example clarifies the script description, the subsequent discussion of another case will put the script to work in a "real-world" archaeological situation drawn from literature. Conclusions will follow.

3. Correspondence Analysis: a short introduction

CA is an exploratory technique that graphically represents the relations among both rows and columns of contingency tables. The visual display of data helps the interpretation and allows patterns to emerge. The technique displays both rows and columns in a reduced-dimensional space by decomposing the total inertia (i.e., the variability) of the data table and identifying the factors that best synthesize the data variability. The graphical output of CA is a set of two-dimensional scatterplots where rows and/or columns are represented as points. The factors may be sorted in order of decreasing amount of inertia summarized: the first one summarizes the highest amount, while the second will be associated with the second largest proportion, and so on.

On the scatterplot, the distance between data points of the same type (i.e., row-to-row) is related to the degree to which the rows have similar profiles (i.e., relative frequencies of column categories). The same applies for the column-to-column distance. The more points belonging to the same set are close to each other, the more similar their profiles are. The origin of the axes represents the centroid (i.e., the average profile, corresponding the table's marginal profile), and can be conceptualized as the "place" where there is no difference between profiles or, more formally (and to recall the chi-square terminology), it represents the null hypothesis of homogeneity of the profiles (Greenacre 2007, 32). The more different are the latter, the more the points will be spread on the plane away from the centroid.

In concluding this introductory section, a few words have to be said about outliers. Outliers are row/column profiles that dramatically deviate from the others, for example, for having very small frequencies. These affect the graphical output of CA, by setting far away from the centroid. As stressed by Baxter, Cool (2010, 220), outliers may dominate the plot, and cause the other profile points to cluster together. Greenacre (2011) shows an alternative representation that adjusts the visualization (i.e., not influenced by outliers), so that the interpretation of plots may be assured even in presence of outliers. This can be accomplished by using the Greenacre's Standard Biplot (Greenacre 2007, 101-102; also called Contribution Biplot in Greenacre 2011, 9) that can be easily obtained by the package "ca" via the R script here described.

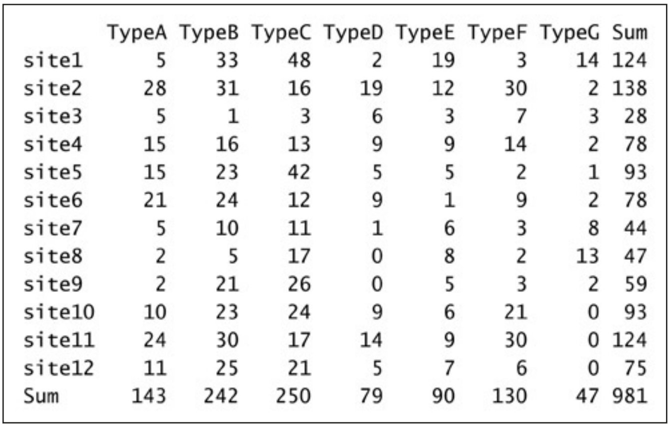
4. R script for Correspondence Analysis: worked examples

4.1 *Requirements to run the script*

In order to run the script, there are just two basic requirements. The first is to install the required R packages (along with all their dependencies), namely "ca" and "FactoMineR" (see § 1 for references). Should the user not know how to install them, the apposite commands can be found right in the first lines of the script. Users can copy and paste them into the console to have R automatically perform the installation. The second requirement is to have some data to feed into R. The contingency table must be saved as tab-delimited text file (.txt). This can be easily accomplished in any statistical program (even in Microsoft Excel).

To enter the data into R is straightforward. Upon running the script, a window will pop-up prompting the user to select the source data file. The script will then produce, along with graphical outputs, a textual one saved in the program's working directory as a text file (named "output_CorrespondenceAnalysis.txt"). It contains the CA output statistics relevant to the interpretation of the data: e.g., original contingency table, row/column profiles, association coefficients, chi-square test, CA principal inertias, rows/columns coordinates, etc.

Once the dataset has been fed into R, the CA will be run automatically and users can just wait for the program to perform the needed steps. Upon completion (which should take just matter of seconds, depending on the computer's speed), the user will have on the screen a series of windows displaying the analysis' results that will be described in the following, as well as the aforementioned textual summary. Admittedly, there is little room for interaction with users during the main body of the analysis. I wish to make clear that, while someone could consider this a flaw, in my opinion it perfectly



```
        TypeA TypeB TypeC TypeD TypeE TypeF TypeG Sum
site1       5    33    48     2    19     3    14 124
site2      28    31    16    19    12    30     2 138
site3       5     1     3     6     3     7     3  28
site4      15    16    13     9     9    14     2  78
site5      15    23    42     5     5     2     1  93
site6      21    24    12     9     1     9     2  78
site7       5    10    11     1     6     3     8  44
site8       2     5    17     0     8     2    13  47
site9       2    21    26     0     5     3     2  59
site10     10    23    24     9     6    21     0  93
site11     24    30    17    14     9    30     0 124
site12     11    25    21     5     7     6     0  75
Sum       143   242   250    79    90   130    47 981
```

Tab. 1 – Table displaying the frequency of 7 pottery types across 12 sites, as input to the R script.

complies with the logic of a script intended for inexperienced and average users, who can just wait for the analysis to be performed. More experienced users, as already stressed, have already the expertise to use the script (or parts thereof) in a more personal way.

In the following paragraphs, the analytical output of the script will be described. The description will be interspersed with a discussion of the context of use of the information provided by the script. Emphasis will be put on the graphical display of the script, and reference will be made to the textual output when relevant to the discussion.

4.2 WORKED (fiCTIONAL) EXAM P LE

For illustrative purposes, a contingency table is created with 12 rows and 7 columns (Tab. 1). It represents the fictional distribution of seven pottery types across twelve sites. The analyst’s interest could lie in understanding if a correspondence exists between sites and pottery types; in other words, whether types are evenly distributed across sites, or if a pattern of association exists between sites and types. This will be accomplished by means of CA.

4.2.1 Association between rows and columns

The preliminary interest could be in the strength of association between rows and columns of the table. This information is provided by a bar chart (Fig. 1a). It shows the magnitude of the correlation coefficient on the right side, compared with the overall range (0.0-1.0) of the coefficient (on the left). A reference line indicates the threshold (0.20) above which the correlation can be considered important (Bendixen 1995, 576; Healey 2013, 289-290). It should be noted that the correlation coefficient is the square root of the table's inertia, and it turns out to correspond to the *phi* coefficient used to measure the strength of association between two categorical variables (Greenacre 2007, 28, 61). In our example, the correlation coefficient is equal to 0.57 pointing to a strong association (Healey 2013, 289). It should be also noted that the existence of a significant dependence between rows and columns could be tested via the chi-square test (on this test see Cool, Baxter 2005; Drennan 2009, 182-188). Should the user be interested in it, the textual output of the script provides the results of the test, which in our case turns out to be significant (chi-square: 319.92; df: 66; *p*: < 0.001).

4.2.2 Number of dimensions useful for data interpretation

Before delving into the core of CA results, the user should decide how many dimensions could be considered relevant for the interpretation of the data. In other words, how many axes he must take into consideration in order to have a good representation of the patterns of association.

It has to be stressed that this is one of the "thorniest" problem (Preacher *et al.* 2013, 29) affecting CA as well as Factor Analysis, Principal Components Analysis (PCA) and Multidimensional Scaling (see, e.g., Jackson 1993; Wilson, Cooper 2008; Van Pool, Leonard 2011, 296-299). As stressed by Hair *et al.* (2009, 591), in selecting the optimal number of dimensions the analyst is faced with the need of a trade-off between the increasing explained data variability deriving by keeping many dimensions versus the increasing complexity that can make difficult the interpretation of more than two dimensions. Like for other techniques, for which a number of approaches exists each having its pros and cons (overviews in Worthington, Whittaker 2006, 820-822; Wilson, Cooper 2008), also in CA there is no clear-cut rule guiding the analyst's choice (Lorenzo-Seva 2011, 97) and different approaches have been proposed.

A more informal approach leans toward considering the number of useful dimensions fixed by the very analyst's ability to give meaningful interpretation of the retained axes (Benzécri 1992, 398; Blasius, Greenacre 1998, 25; Yelland 2010, 13). In other words, dimensions that cannot be sensibly interpreted can be considered the result of random fluctuations among the residuals (Clausen 1998, 25).

Another approach would be to keep as many dimensions as necessary to account for the majority of the total inertia, setting a cut-off threshold at an arbitrary level, say 90% (see, in the context of Factor Analysis, Van Pool, Leonard 2011, 296). On the other hand, Hair *et al.* (2009, 591) suggest that dimensions whose inertia is greater than 0.2 (in terms of eigenvalue) should be included in the analysis.

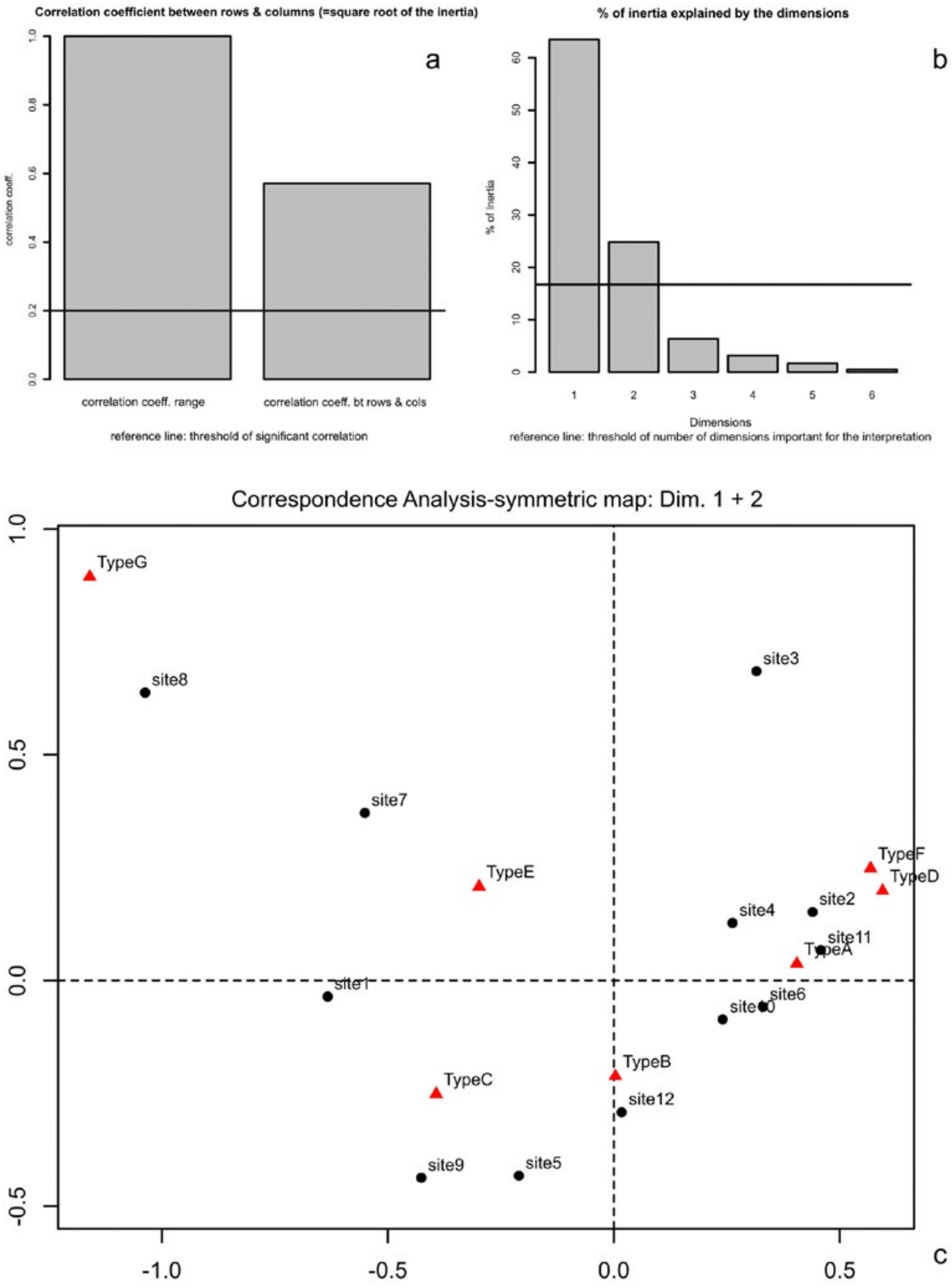


Fig. 1 – CA on data of Tab. 1. Charts provided by the R script. a) Bar chart showing the coefficient (right) for the correlation between rows and columns of the table. A reference line indicates the threshold of “significant” correlation. b) Bar chart showing the percentages of inertia explained by the CA dimensions. A reference line suggests the threshold above which a dimension should be considered important for data interpretation according to the average rule. c) Symmetric map of CA on Tab. 1, showing the first 2 dimensions (from the “ca” package).

Another frequently used method is the inspection of the scree plot, adapted from the context of PCA (Cattell 1966). Dimensions are plotted in order of the decreasing amount of explained inertia, resulting in a falling curve. The point at which the latter shows a bend (so called "elbow") can be considered as indicating an optimal dimensionality (e.g., Clausen 1998, 24; Drennan 2009, 286-288; Van Pool, Leonard 2011, 296-297). It is worthy of note that this method has been found to perform fairly well (Zwick, Velicer 1986, 440; Bandalos, Boehm-Kaufman 2009, 81; Lorenzo-Seva 2011, 97).

The average rule, as termed by Lorenzo-Seva (2011, 97), is yet another method, which is equivalent to the Kaiser's rule in the context of PCA (Wilson, Cooper 2008, with references). According to this rule, analysts should retain all the dimensions that explain more than the average inertia (expressed in terms of percentages), the latter being equal to 100 divided by the number of dimensions (i.e., the number of rows or columns, whichever is smaller, minus 1). Unfortunately, in the context of PCA, this method seems to overestimate the dimensionality of the solution (Wilson, Cooper 2008, 866; Lorenzo-Seva 2011, 97).

Saporta (2006, 209-210) has suggested the use of the Malinvaud's test as guidance for the dimensionality of the CA solution (see also Camiz, Gomes 2013a, 12). In practice, referring to Saporta's book or Camiz-Gomez's article for the computational details (see also Rakotomalala 2013, 7), this sequential test checks the significance of the remaining dimensions once the first k ones have been selected. As stressed by Saporta himself and empirically tested by Rakotomalala (2013), it seems to overestimate the number of dimensions as the table's grand total increases.

Finally, Lorenzo-Seva (2011) has interestingly adapted to CA a method developed for PCA, called Parallel Analysis. Its rationale is that nontrivial dimensions should explain a larger percentage of inertia than the dimensions derived from random data. While this method outperforms the aforementioned average rule, it seems to suggest a dimensionality of the solution comparable to the one that can be derived from the scree plot, at least in the illustrative example discussed by the scholar (Lorenzo-Seva 2011, 101, fig. 1).

In front of the sizable number of different approaches, each one having its pros and cons, I would lean toward a middle ground as to the problem of the dimensionality of the CA solution, trying to conciliate formal testing, on the one hand, with conceptual interpretability as dimension-retention criterion, on the other hand. I would agree with Worthington, Whittaker (2006, 822) who lucidly state that «in the end, researchers should retain a factor only if they can interpret it in a meaningful way no matter how solid the evidence for its retention». In their opinion, exploratory approaches are «ultimately a combination of empirical and subjective approaches to data analysis because the job is not complete until the solution makes sense». Within this general framework, I would also agree with Bandalos, Boehm-Kaufman (2009, 80-81) as to the need to compare and find a balance between different methods, provided that one deals with significant factors according to the chi-square statistic. The methods provided by the script are the average rule, the scree plot, and the Malinvaud's test. A fourth criterion, namely the retention of dimensions whose eigenvalue is greater than 0.2 (*sensu* Hair *et al.* 2009 previously quoted), can be easily put to work thanks to the script output (via the "ca" package), as will be described shortly. The reason for the choice of these methods rests on the opportunity to provide the users with the possibility to compare at least four of the criteria previously illustrated. Lorenzo-Seva's Parallel Analysis as applied to CA would be interesting to implement, but, admittedly, is beyond my current programming expertise.

As for the average rule in the context of our worked example, any axis contributing more than the average percentage of inertia (100/11=9% in terms of rows, 100/6=16.7% in term of columns) should be considered important for the interpretation of the data (see, e.g., Bendixen 1995, 577). It must be acknowledged, however, that interesting patterns can emerge by inspecting more than just the first two dimensions, as rightly stressed by Baxter (1994, 120). With this warning in mind, the bar chart provided by the script can be used as a guidance in the choice of the relevant dimensions (Fig. 1b). Dimensions are plotted in order of the decreasing amount of explained inertia. A reference line represents the threshold above which a dimension can be considered important according to the average rule. In our case, a 2-dimensional solution seems appropriate, with the first explaining over 60% of the inertia, and the second about 20%. It must be noted that the threshold represented by the reference line is also indicated in a specific section of the script's textual output.

The same chart can be read off as a scree plot: the point at which a bend is evident in the falling curve described by the histograms can be taken as indicating *an* (not *the*) optimal dimension. It is worth noting that the number of dimensions suggested by the chart, once it is read off as a scree plot, is consistent with the dimensionality suggested by the average rule.

The result of the Malinvaud's test is reported in a specific section of the script's textual output. In our case, only the first three dimensions seems to be important since their p value is below 0.05, while the other three have a value equal to 0.167, 0.573 and 0.825 respectively.

As far as the greater-than-0.2 rule is concerned, the dimensions complying with that criterion can be located by inspecting the script's textual output, which reports the list of dimensions with associated eigenvalues (after the "ca" package). According to this rule, only the first dimension, accounting for more than half of the total inertia (i.e., 63.5%), should be retained.

The difference between the four methods underscores the need to compare and find a balance between multiple dimension-retention criteria. In our case, a 2 or 3-dimensional solution seems appropriate.

4.2.3 Interpreting the CA scatterplot: dimensions interpretation

The script provides the symmetric plots (rows and columns; rows only, columns only) for the significant dimensions. The plots are obtained from the "ca" package. The plot, showing the row (point) and column (triangle) profile points at the same time, is reproduced in Fig. 1c. To interpret the plot it is useful to clarify that we are interested in interpreting the relative position of the row points in the space defined by the columns. In other words, we seek to understand the similarity of the sites on the basis of the proportion of pottery types present in each location. The next step for the interpretation is to assess what column category (i.e., type) is actually determining the dimensions. In a sense, we are going to give "names" to the dimensions.

To accomplish this, two ways are made available by the script. The first, which is not natively available from any of the packages taken into account here and has required some additional programming, is to inspect the bar plot in Fig. 2a. The contribution (in permills) of pottery types to the definition of the first four dimensions is displayed. The reference line indicates the threshold (average contribution) above which any contribution has to be considered important for the definition of that dimension (Greenacre 2007, 82). It can be seen that type C, F, and G have a major role in the definition of the first dimension, with the first and last of the three also having a large contribution to the second dimension. Incidentally, it must be noted that type A and B have a large contribution only to the third and fourth dimension. Should the analyst be interested in those pottery types, he can proceed to inspect the other plots provided by the script, representing the first dimension and the third or the fourth one.

The second option is the Standard Biplot (also called Contribution Biplot; see § 3) provided by the "ca" package (Fig. 2b). In this plot, while the position of the row profile points is unchanged relative to that in Fig. 1c, the distances of the column points from the centroid are related to the contribution that each column category gives the principal axes (Greenacre 2007, 101-103, 268-270). Besides, the closer an arrow is (in terms of angular distance) to an axis (or to a pole thereof) the greater is the contribution of the column category on that axis relative to the other axis. If the arrow is halfway between the two, its column category contributes to the two axes to the same extent. It is evident that type F has a major contribution to the positive pole of the first dimension, while type C and G have a major contribution to the definition of both the first and second dimension. In this respect, while C and G both contribute to the negative pole of the first dimension, they contribute

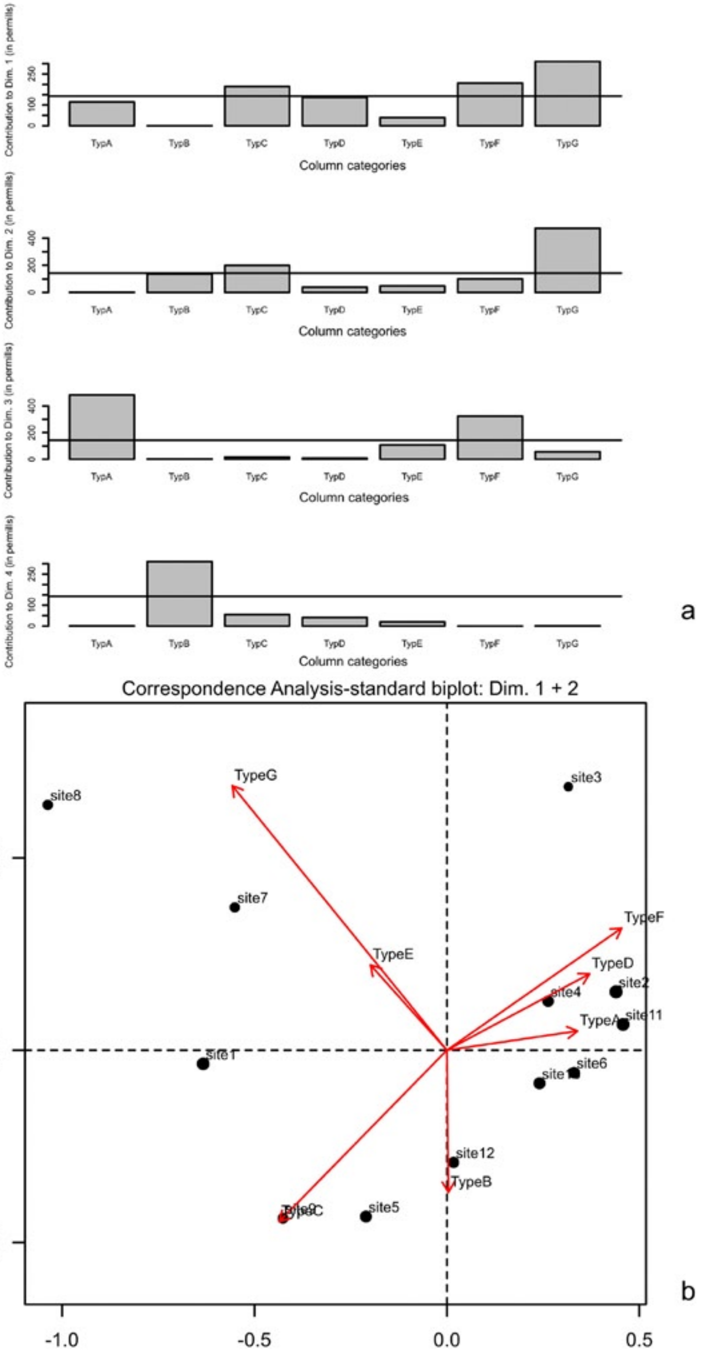


Fig. 2 – CA on data of Tab. 1. Charts provided by the R script: interpretation of the dimensions. a) Bar chart showing the contributions (in permills) of column categories to the first 4 dimensions; a reference line indicates the average contribution. b) Standard Biplot showing the first 2 dimensions (from the "ca" package). Note: the length of each arrow joining the column points to the origin is related to the contribution that each column category makes to the principal axes.

to different poles of the second dimension, the negative and the positive one respectively. The information provided by Figs. 2a and 2b complement each other and both could be used in reports or publications.

In fact, while in Fig. 2a it is possible to see which category has a major contribution to the dimensions, from Fig. 2b we can gain the same information directly in the context of the scatterplot and in addition we can have an idea of what pole of the dimensions the column categories are actually contributing to. Indeed, the Standard Biplot has another interesting interpretative potential that will be touched upon later on (§ 4.2.6)

4.2.4 Interpreting the CA scatterplot (continued): correlation between row profiles and dimensions

The next step is to consider the correlation between the row profiles and the dimensions. This means that, after having given "names" to the dimensions (i.e., after having located what column category has determined the dimensions), we can understand how row categories (sites, in our example) relate to the dimensions. This can be done by inspecting the bar plot (not natively returned by any package) provided by the script (Fig. 3), where the correlation (ranging from 0.0 to 1.0) between row categories and the dimensions is displayed.

The reader is referred to Greenacre (2007, 86) for a full coverage of the way in which CA computes these figures. It suffices here to stress that almost all the sites (i.e. row categories) have a strong correlation with the first dimension, with the exception of site 3, 5, and 12. These have a strong correlation with the second dimension instead. The analyst may refer back to the scatterplot (Fig. 1c) to have an idea to which pole of the dimensions these correlations refer.

4.2.5 Quality of the representation

Finally, the analyst has to take into consideration the fact that not all the points could be well displayed in the chosen dimensions. To assess the quality of the display, he can consult the statistics provided by the "ca" package showing both on the R console and in the textual output of the script, or inspect the bar chart provided by the script itself (Fig. 4). It can be seen that almost all the sites are well displayed by the first two dimensions or, in other words, these dimensions explain the greatest percentage of the inertia of those profiles. Only site 6 and 10 turn out to be poorly displayed, implying that the position of those two points on the scatterplot must be evaluated with caution.

4.2.6 Assembling the whole picture

From the preceding guidelines it should be apparent that by means of CA we may have a clearer and richer picture of the patterns of association

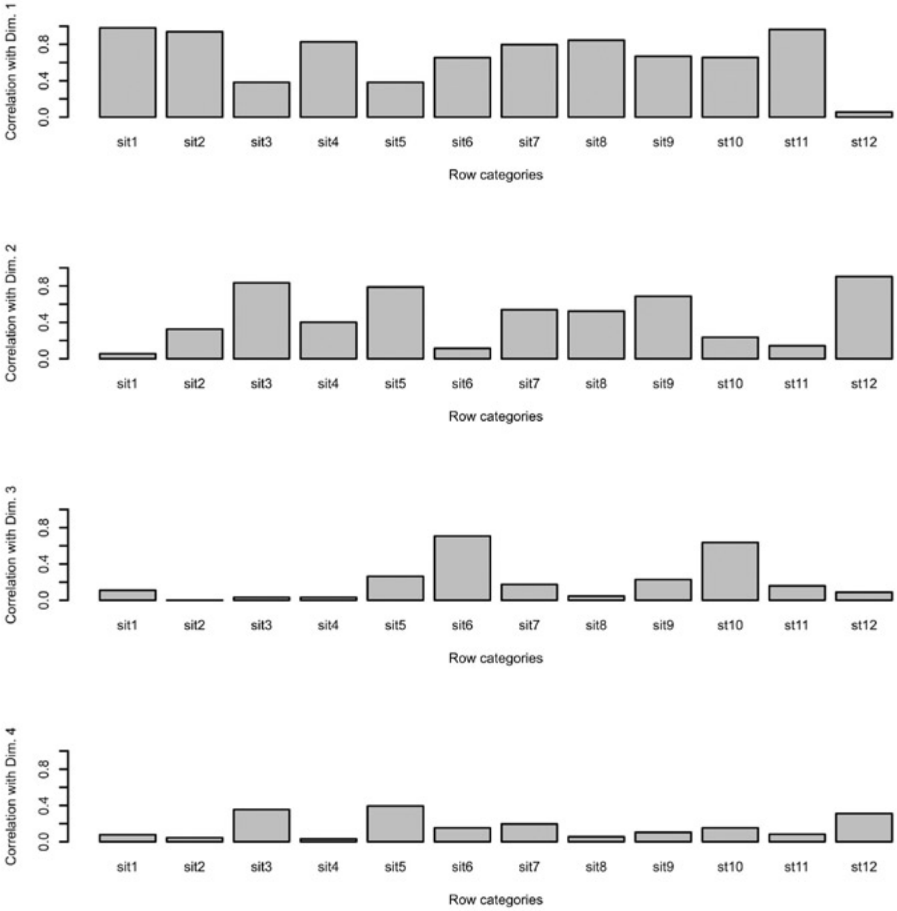
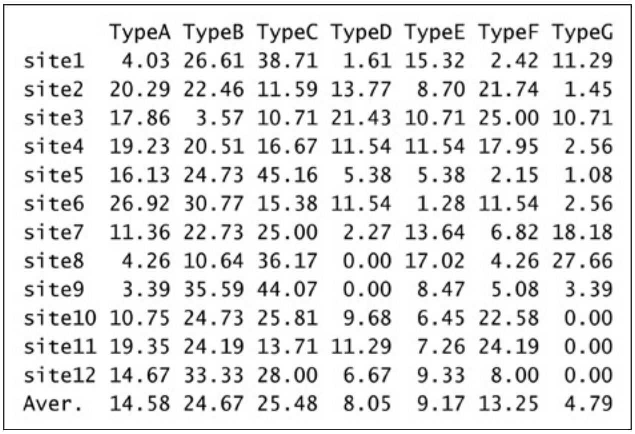


Fig. 3 – CA on data of Tab. 1. Charts provided by the R script. Bar chart showing the correlation between the row categories (sites) and the CA dimensions.

between sites and type and, more importantly, we can dissect patterns of variations encoded in our data. CA allowed the isolation of two main trends (i.e., dimensions) of variation in our dataset, with the first being far more important in that it accounts for more than half the total data variability (i.e., inertia). The first two dimensions together explain almost 90% of the inertia (actually, 88.3%).

It has been possible to assess that the first dimension is determined by the opposition between type F (positive pole), on the one hand, and C and G (negative pole) on the other. The second dimension (accounting for a lesser



```
       TypeA TypeB TypeC TypeD TypeE TypeF TypeG
site1   4.03 26.61 38.71  1.61 15.32  2.42 11.29
site2  20.29 22.46 11.59 13.77  8.70 21.74  1.45
site3  17.86  3.57 10.71 21.43 10.71 25.00 10.71
site4  19.23 20.51 16.67 11.54 11.54 17.95  2.56
site5  16.13 24.73 45.16  5.38  5.38  2.15  1.08
site6  26.92 30.77 15.38 11.54  1.28 11.54  2.56
site7  11.36 22.73 25.00  2.27 13.64  6.82 18.18
site8   4.26 10.64 36.17  0.00 17.02  4.26 27.66
site9   3.39 35.59 44.07  0.00  8.47  5.08  3.39
site10 10.75 24.73 25.81  9.68  6.45 22.58  0.00
site11 19.35 24.19 13.71 11.29  7.26 24.19  0.00
site12 14.67 33.33 28.00  6.67  9.33  8.00  0.00
Aver.  14.58 24.67 25.48  8.05  9.17 13.25  4.79
```

Tab. 2 – Row profiles of Tab. 1; the average row profile is also shown.

amount of variability) is determined by the opposition between G (positive pole) and C (negative pole). It is now possible to interpret the position of the sites relative to the dimensions in terms of the different influence of each dimension (i.e., pottery types) on the sites. The more they lie on the right (the positive side of the first dimension) the more they will be "associated" with type F or, put another way, the more type F will make a high proportion in their assemblages. This does not mean that sites on that side of the plot will not have type A and D. It does mean, however, that the proportion of type F will be greater than one of the other two types. The more the sites will lie to the left (negative pole of the first dimension), the more they will be "associated" with types G and C. Moreover, with respect to the second dimension, the more the sites lie in the upper part the plot, the more they will be correlated to type G, while type C will make a higher proportion in the assemblages of the sites lying in the lower part of the plot.

As seen in the bar chart in Fig. 3, site 2, 4, 10, and 11 have a high correlation with the first dimension (i.e., type F). It is possible to take a look at the table of row profiles (Tab. 2) to see that in those sites a higher-than-average proportion of type F is present. The only exception is site 6, which is displayed near the previous four site points even if that pottery type makes a proportion lower than the average. The reason is that site 6 is not well displayed by the first two dimensions, as seen in Fig. 4. As for the other sites, 5, 9, and 12 have a high correlation with the negative pole of the second dimension (i.e., type C) and, accordingly, show a higher-than-average proportion of that particular type. Finally, site 1, 7, and 8 are highly correlated with the first (negative

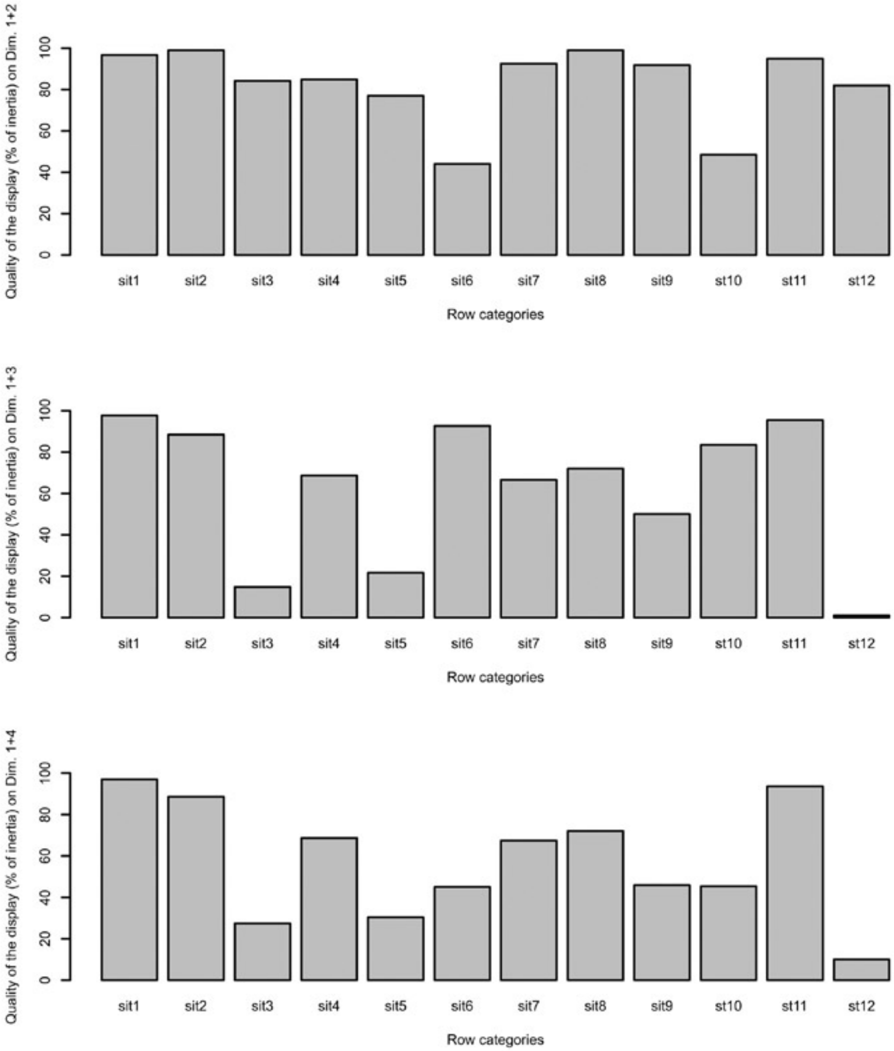


Fig. 4 – CA on data of Tab. 1. Charts provided by the R script. Bar chart showing the quality of the representation of row categories (sites) on the dimensions 1+2, 1+3, and 1+4.

pole) and second (positive pole), both determined by type G, which makes a higher proportion on these sites.

It has to be noted that the Standard Biplot (Fig. 2b) also gives an idea of the relative frequency of a given pottery type in the sites' assemblages. This is one of the advantages referred to in § 4.2.3. For example, consider the imaginary axis to which the arrow representing the type G belongs, and let us line up on it the projections of the row profile points. The profile points whose projection intersects the axis on the same side of the arrow are those having a higher-than-average proportion of that pottery type. Those intersecting the axis on the opposite side are those having a lower-than-average proportion. In addition, the more a projection intersects the axis away from the centroid, the greater will be the difference between the average and the proportion that the pottery type makes on the profiles (Greenacre 2007, 103). For example, taking into account sites 1, 7, and 8, the one whose projection lies further from the origin is site 8, which has the highest proportion of that type (27.66%). The second and third are, respectively, site 7 (18.18%) and 1 (11.29%).

The second advantage of the Standard Biplot comes into play in the presence of outliers (§ 3 with references therein). Should outliers be present, since generally they are profiles with a low contribution to the inertia, the Standard Biplot provides the possibility to reduce the distortion in the graphical display (i.e., plotting the outliers too far from the centroid). In fact, in this plot, the smaller the contribution of a category to the definition of the dimensions, the more it will be pulled in toward the centroid.

4.2.7 Extension: clustering rows and/or columns

The script provides the facility to perform a cluster analysis (Baxter 1994, 140-184; Shennan 1997, 216-264; Drennan 2009, 309-320) over the CA results. This is accomplished via the "FactoMineR" package. Often the user could be interested in isolating clusters of points on the CA scatterplot (see, e.g., Wallin 2010, 70). To keep with our example, he could be willing to indicate on the scatterplot groups of sites that are similar in terms of their assemblage profiles. This could be accomplished in an informal way, grouping "by eye" the points lying one near the other on the plot. Indeed users may require a more formal method.

Greenacre (1988; 2007, 113-120) describes a method particularly well suited to the underlying logic of CA (Greenacre 1988, 41), whose algorithm can be described as follows (Greenacre 2007, 116): rows are progressively aggregated in a way in which every successive merging produces the smallest change in the table's inertia, and this process goes on until the table is reduced to just one row "consisting of the marginal columns of the original table" (Greenacre 2007, 116, 117, fig. 15.4). The same applies to columns. The underlying logic lies in the fact that rows (or columns) whose merging produces a small change in table's inertia have similar profiles. This procedure can be thought of as maximizing the between-group inertia and minimizing the within-group inertia (Greenacre 2007, 116). The successive merging of rows (or columns) can be graphically depicted as a dendrogram. Each level

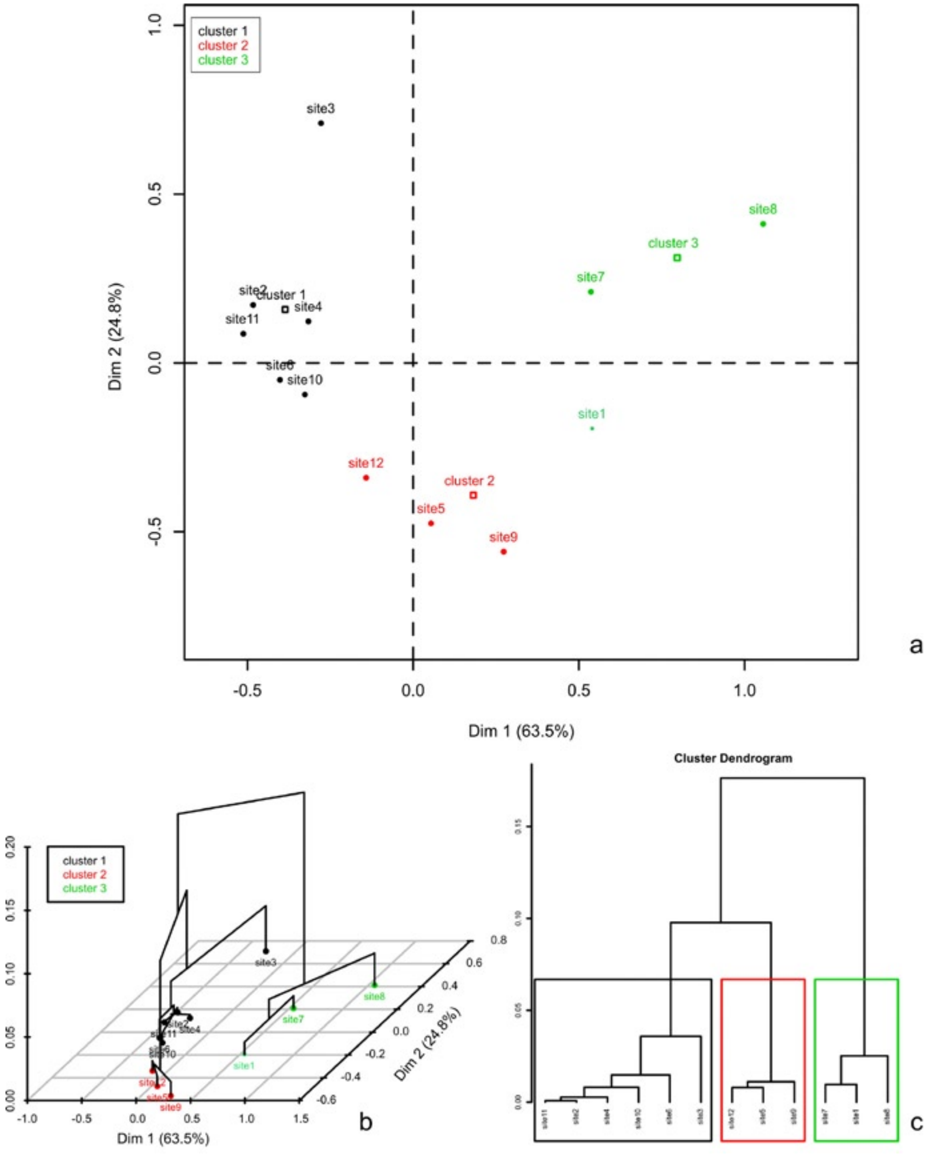


Fig. 5 – CA on data of Tab. 1. Charts provided by the R script. a) CA map showing row categories (sites) with different colours indicating different clusters. b) As previous figure, with cluster tree displayed on the map. c) Cluster tree with optimal cluster partition (boxes) as suggested by "FactoMineR". All charts from the "FactoMineR" package.

at which the merging occurs corresponds to the associated reduction of the table's inertia.

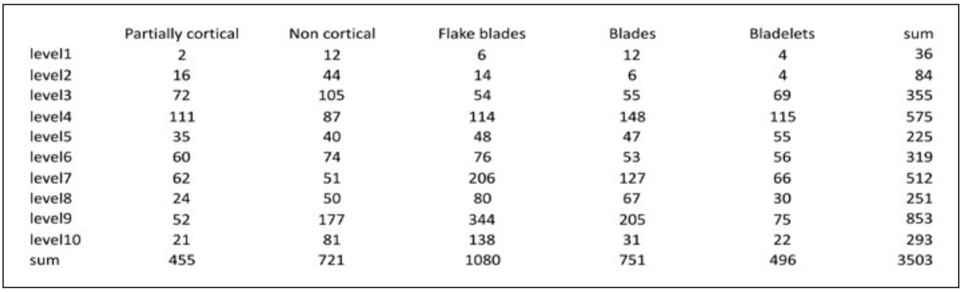
A method essentially similar is that provided by the "FactoMineR" package (Lê *et al.* 2008; Husson *et al.* 2010; 2011, 177-185) used by the script. It returns three plots: the first is the CA scatterplot with points (in our example, row profile points) coloured on the basis of the clusters they belong to (Fig. 5a); the second is quite similar, with the clusters tree plotted directly onto the scatterplot (Fig. 5b); the third is the clusters tree with the optimal level of division indicated by coloured boxes (Fig. 5c).

Even though different approaches exist in cluster analysis to decide how many clusters can be read out of the results (i.e., cutting the dendrogram at a particular height), ranging from informal (Drennan 2009, 316) to formal ones (so-called "stopping rules"; see e.g. Milligan, Cooper 1985, 163-167; overview in Baxter 1994, 161-165; Everitt *et al.* 2011, 95-96), and acknowledging the fact that the problem is a «difficult one for which no completely satisfactory solution exists» (Baxter 1994, 162), "FactoMineR" natively suggests an optimal partition. While its mathematical details are beyond the scope of this article (and I refer the reader to the references provided), suffices here to say that, as made clear by Husson *et al.* (2011, 185), a division into Q (i.e., a given number of) clusters is suggested when the increase in between-group inertia attained when passing from a Q-1 to a Q partition is greater than that from a Q to a Q+1 clusters partition. In other words, during the process of rows (or columns) merging, if the following aggregation raises highly the within-group inertia, it means that at the further step very different profiles are being aggregated.

To keep with our fictional example, this means that the sites belonging to the same cluster (2, 3, 4, 6, 10, 11; 1, 5, 9, 12; 7, 8) are those with more similar profiles. Referring back to the original contingency table, those rows could be collapsed into two distinct groups, and this would produce the least decrease in the table's inertia since, as said, the sites belonging to those two groups have the more similar profiles in terms of pottery types. This could be relevant for the sake of any further archaeological interpretation since it could provide the bases to hypothesize, for instance, that the sites could represent two different chronological horizons, or could belong to two different cultural traditions, and so forth.

4.3 *Additional example*

In this paragraph it is considered the dataset illustrated, but not further discussed, by Shennan (1997, 355-357). It is made up of 10 rows and 5 columns (grand total: 3503) and concerns the counts of different lithic types from ten levels at the Palaeolithic cave at Ksar Akil (Lebanon) (Tab. 3). An attempt will be made to use CA as a mean to reveal patterns



	Partially cortical	Non cortical	Flake blades	Blades	Bladelets	sum
level1	2	12	6	12	4	36
level2	16	44	14	6	4	84
level3	72	105	54	55	69	355
level4	111	87	114	148	115	575
level5	35	40	48	47	55	225
level6	60	74	76	53	56	319
level7	62	51	206	127	66	512
level8	24	50	80	67	30	251
level9	52	177	344	205	75	853
level10	21	81	138	31	22	293
sum	455	721	1080	751	496	3503

Tab. 3 – Cross-tabulation of the frequency of five lithic types across ten levels from the Palaeolithic cave at Ksar Akil (Lebanon) (after Shennan 1997).

of assemblage variation and to understand how lithic types relate to any pattern pinpointed.

The result of the chi-square test is significant (chi-square: 448.6, df: 36, *p*: <0.0001) and the square root of the inertia is 0.36. While different views exist on how to characterize the correlation strength, since scales and their boundaries are often defined using subjective arguments varying from a research field to another, 0.36 can be thought of as pointing to the existence of an association between row and column categories that someone could label as low (Rowntree 2000, 170) or moderate (Taylor 1990, 37), while other could interpret it as moderately large (Cohen, Lea 2004, 211) or even strong (Healey 2013, 289 table 11.12). Taking a middle ground, we could define our correlation as moderate.

While both the first and second dimensions are below the 0.2 threshold (according to the aforementioned criterion as suggested by Hair *et al.* 2009), the average rule and the scree plot point to a 2-dimensional solution. As for the Malinvaud's test, the first three dimensions have a p value well below 0.01, while the p value of the fourth dimensions is 0.148. It is worthy of note that the first two CA dimensions capture the majority of the data variability (59.4% and 32.0% respectively), for a total of 91.4% (Fig 6a). All the categories are therefore well represented on that plane, i.e. have good quality of the display (Fig. 6b).

Since the interest lies in understanding the composition of the levels as far as the proportion of different lithic types is concerned, the interpretation of the CA results will be centered on the examination of the position of the level points in the space defined by the lithic type categories.

The symmetric biplot clearly depicts a major division between two broad groups lying on the opposite poles of the first (horizontal) dimension (Fig. 7). This holds true both for the levels (i.e., rows) and lithic types (i.e., columns). By inspecting the Standard Biplot (Fig. 8), on the one hand, and the bar chart

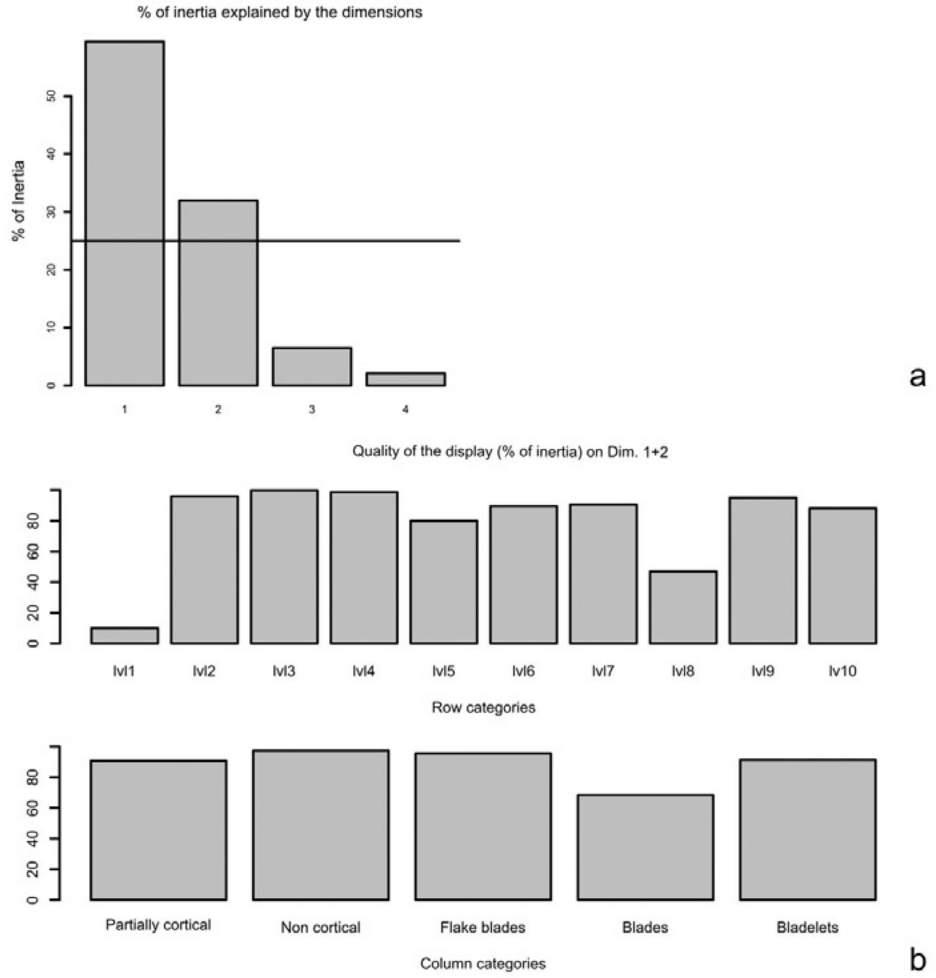


Fig. 6 – CA on data of Tab. 3. a) Percentage of inertia explained by the dimensions. Reference line: threshold above which a dimension should be considered important for data interpretation according to the average rule. b) Histograms showing the quality of the display of row and column categories on the first two CA dimensions.

of the contribution of column categories to the principal dimensions (Fig. 9a), on the other hand, it is apparent that the types "partially cortical" and "flake blades" are actually defining the opposite poles of the first dimension. The "non cortical" and "blades" categories are the major contributors to the definition of the second dimension. Levels 7, 8, 9 and 10 have a high correlation with

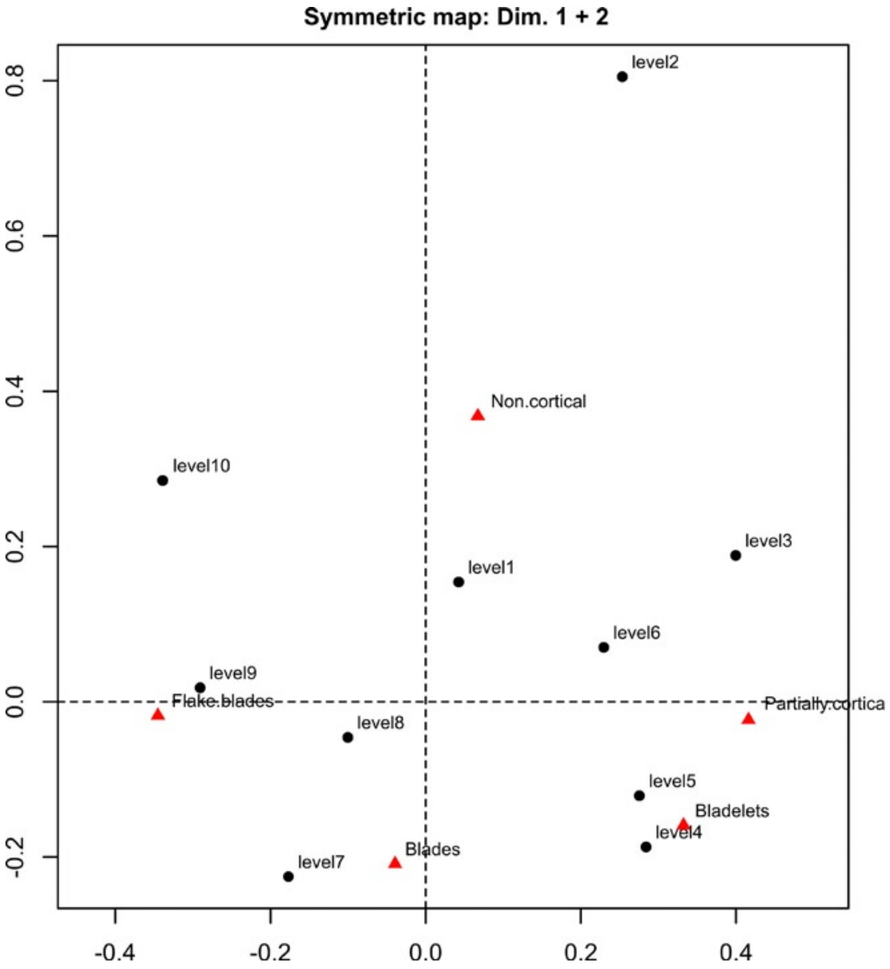


Fig. 7 – Symmetric map of CA on Tab. 3, showing the first 2 dimensions.

the "flake blades" (i.e., negative pole of first dimension) (Fig. 9b), while levels 3, 4, 5, and 6 are mainly correlated with "partially cortical" (i.e., positive pole of the same dimension). Moreover, level 2 has the highest correlation with "non cortical" type (positive pole of the second dimension), and level 7 is also correlated with "blades" (negative pole of the same dimension).

On the whole, it seems that in the data a trend can be discerned pointing to a shift from the "partially cortical" category featuring the upper (i.e., later) levels to the "flake blades" category having a major proportion in lower (i.e., earlier) levels. It is worth noting that this picture is consistent with the

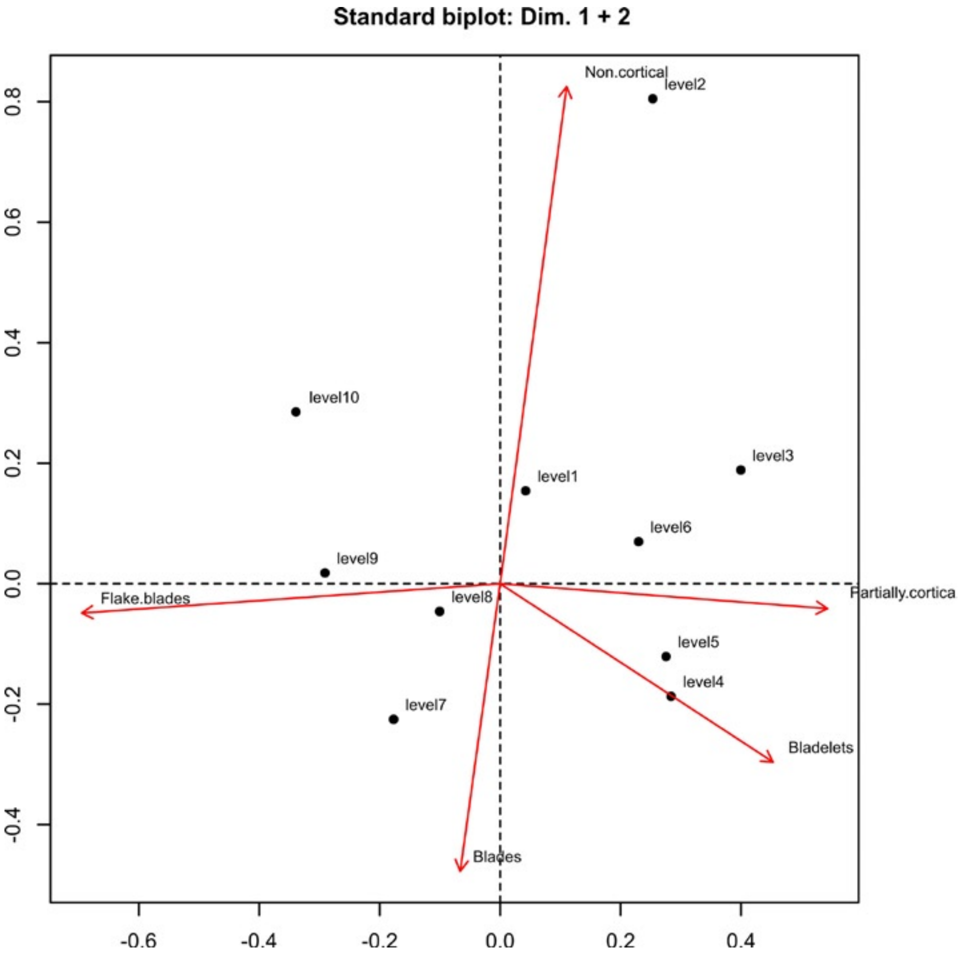


Fig. 8 – Standard Biplot of CA on Tab. 3, showing the first two dimensions. For the interpretation of this type of Biplot, see caption of Fig. 2.

remarks of Goring-Morris, Bergman (1987) who located in levels 8 and 6 (lying on the opposite poles of the first dimension in the present analysis) two major shifts in the development of the lithic production technology at the site. There are grounds to believe that the first CA dimension is successfully capturing the temporal shift from earlier to later levels and related lithic assemblage composition. A slight parabolic curve can be discerned on the CA scatterplot in relation to the spread of level points. It is more apparent by plotting (just for illustrative purposes since, as seen, the first two dimensions

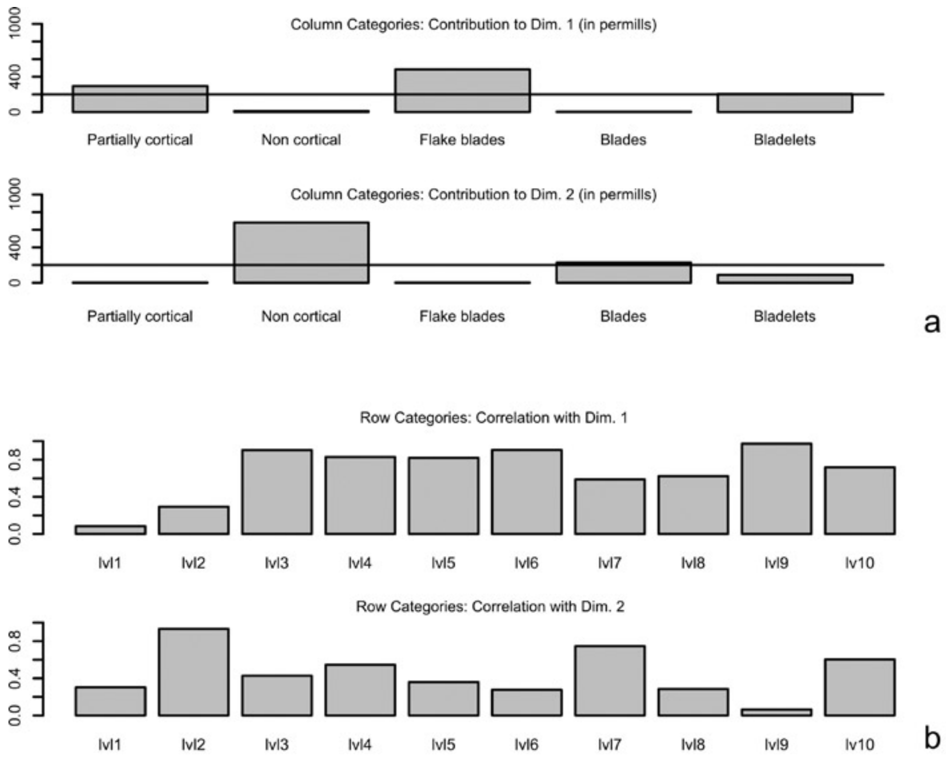


Fig. 9 – a) Histograms showing the contribution (in permills) of the column categories of Tab. 3 (i.e., lithic types) to the definition of the first two CA dimensions. Reference line: average contribution. b) Histograms showing the correlation of row categories (i.e., archaeological levels) with the first two CA dimensions.

account for most of the inertia) the first and third dimensions (Fig. 10). While it should point to the existence of a seriation structure (see references in § 2), it seems too sparse to suggest a good seriation (on this topic, see, e.g., the discussion in Kjeld Jensen, Høilund Nielsen 1997, 43-51). While a seriation is poorly supported by the data, CA's first dimension allows to isolate two major groups of levels, as seen.

If we are interested in a more formal way to locate those groups on the basis of the CA results, the rows hierarchical clustering as provided by the script can be inspected (Fig. 11a-c). It become even more apparent a clear distinction in two major groupings. Again, this is consistent with Goring-Morris, Bergman's (1987, 143) remarks. As a matter of fact, they underscored that the site's sequence is unlikely to represent a developmental continuum. Rather, in their view, groups of levels are likely to represent a

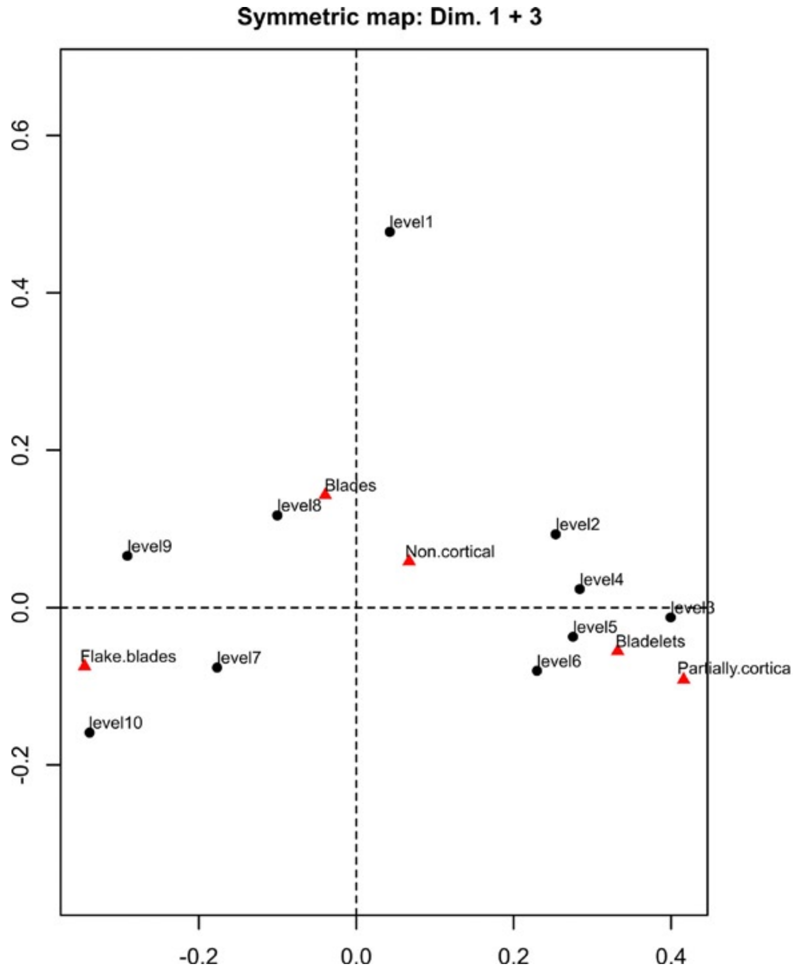


Fig. 10 – Symmetric map of CA on Tab. 3, showing (just for illustrative purposes) the first and third dimensions.

«short-term evolution which is separated from the other groups by breaks in the archaeological sequence». It is apparent that CA seems to support these conclusions. In fact, on the one hand, the very weak seriation structure is consistent with the view of a short-term evolution (interrupted by breaks) as opposed to a developmental continuum. On the other hand, the distinction of two major groups of levels is consistent with the picture of broad chronological horizons representing two main stages of the development of the site's

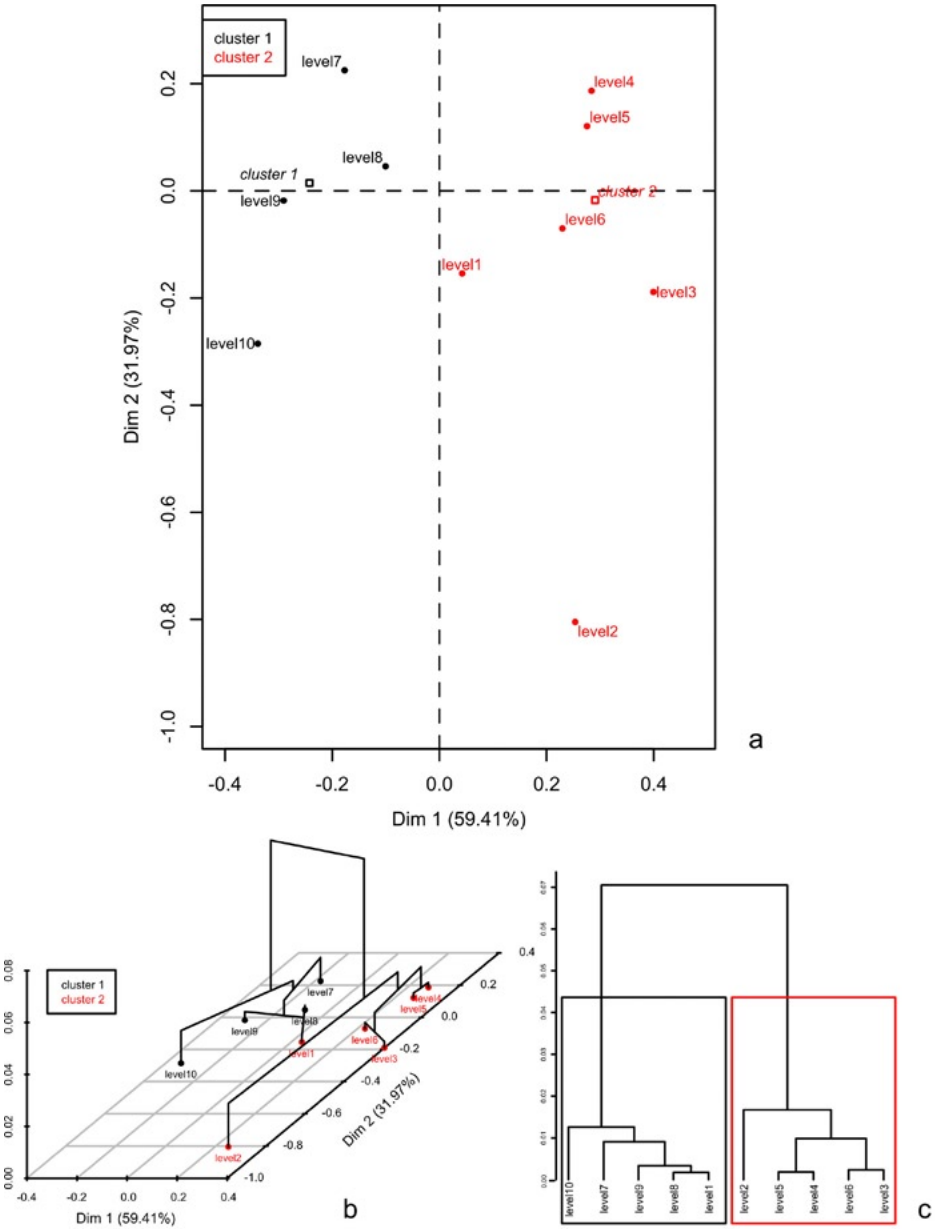


Fig. 11 – Hierarchical clustering of row categories of Tab. 3. a) CA map showing row points (archaeological levels) on the first two dimensions, coloured according to cluster membership. b) CA map with clusters tree superimposed. c) Cluster tree with indication of optimal partition (boxes) in two clusters as suggested by "FactoMineR".

lithic technology. Finally, it should be noted that, since level 1 falls in the same group of the lower levels, the analyst should take into consideration the need to explore further whether or not non-cultural processes (post depositional disturbances or other kinds of events) could have possibly affected the lithic assemblage of that top level.

5. Conclusions

This article has attempted to describe the usefulness of the script written in order to simplify the use of CA in the R statistical environment. While the latter has enormous advantages (free of charge, publication quality graphical outputs, variety of statistical tools available), its command-line structure could intimidate the potential users of CA. This paper has attempted to show how the script can make CA easy to perform in R with just a couple of clicks on the user's part. More importantly, by discussing two worked examples, it has been shown how the script can provide the user with a body of graphical and textual outputs relevant to the interpretation of data structure. The article has also attempted to stress the utility of some graphs not so widely used (to the best of author's knowledge) in the context of archaeological studies (i.e., Greenacre's Standard Biplot), not to mention the utility of the rows/columns clustering as provided by the "FactoMineR" package, and of the Malinvaud's test for the selection of an optimal dimensionality of the CA solution, which has been implemented in R for the first time. Finally, it is hoped that the description of the script will offer a common-sense approach to CA that will prove useful even to the most skeptical user.

Gianmarco Alberti
PhD, University of Udine

Acknowledgements

I wish to thank Michael J. Baxter, Hilary E.M. Cool, and Juan A. Barceló for their criticisms and insightful comments on an earlier draft of this article. I express my sincere thanks to Sergio Camiz whose several critical remarks and insightful and encouraging suggestions allowed me to greatly improve the quality of the article. Thanks are due to him also for stimulating me to implement, with some additional programming effort, the Malinvaud's test. His thoroughly editing of the final version of the manuscript has been also very appreciated. Further, I am grateful to: Julie Josse for her kindness and patience in discussing with me some details of the FactoMineR's clustering algorithm; Michael Greenacre and Urbano Lorenzo-Seva for providing useful bibliographical references on the thorny issue of the selection of the optimal dimensionality of the CA solution; Rico Rakotomalala for promptly replying to my questions about the pros and cons of the Malinvaud's test. Thanks are due to Ray Kidd for the patience and competence shown in correcting the English text of an earlier version of the manuscript. Last but not least, I wish to thank Paola Moscati and Alessandra Caravale for the kindness shown during the whole process of the article's submission and peer reviewing. Needless to say, I am the sole responsible for any error or misunderstanding.

REFERENCES

Alberti G. 2012, *Organizzazione sociale e pratiche comunitarie. Analisi per una ricostruzione del quadro sociale delle comunità eoliane nella Media età del Bronzo*, Unpublished PhD Diss., Department of History and Preservation of the Cultural Heritage, University of Udine.

Alberti G. 2013, *Making sense of contingency tables in archaeology: The aid of Correspondence Analysis to intra-site activity areas research*, «Journal of Data Science», 11, 479-499.

Bandalos D.L., Boehm-Kaufman M.R. 2009, *Four common misconceptions in Exploratory Factor Analysis*, in C.E. Lance, R.J. Vandenberg (eds.), *Statistical and Methodological Myth and Urban Legends*, New York-London, Routledge, 61-87.

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

Baxter M.J., Cool H.E.M. 2010, *Correspondence Analysis in R for archaeologist: An educational account*, «Archeologia e Calcolatori», 21, 211-228.

Bellanger L., Tomassone R., Husi P. 2008, *A statistical approach for dating archaeological contexts*, «Journal of Data Science», 6, 135-154.

Bendixen M. 1995, *Compositional perceptual mapping using chi-squared tree analysis and Correspondence Analysis*, «Journal of Marketing Management», 11, 571-581.

Benzécri J.P. 1992, *Correspondence Analysis Handbook*, New York, Marcel Dekker.

Blasius J., Greenacre M. 1998, *Visualization of Categorical Data*, San Diego-London, Academic Press.

Bølviken E.E., Helskog K., Holm-Olsen I., Solheim L., Bertelsen R. 1982, *Correspondence Analysis: An alternative to Principal Components*, «World Archaeology», 14, 41-60.

Camiz S., Gomes G.C. 2013a, *Joint Correspondence Analysis versus Multiple Correspondence Analysis: A solution to an undetected problem*, in A. Giusti (ed.), *Classification and Data Mining. Studies in Classification, Data Analysis, and Knowledge Organization*, Berlin-Heidelberg, Springer, 11-18.

Camiz S., Gomes G.C. 2013b, *Multiple and joint Correspondence Analysis: testing the true dimension of a study*, «Modulad», 44, 1-21.

Cattell R.B. 1966, *The Scree Test for the number of factors*, «Multivariate Behavioral Research», 1, 245-276.

Clausen S.E. 1998, *Applied Correspondence Analysis. An Introduction*, Thousand Oaks-London-New Delhi, Sage University Press.

Cohen B.H., Lea R.B. 2004, *Essentials of Statistics for the Social and Behavioural Sciences*, Hoboken, Wiley.

Cool H.E.M., Baxter M.J. 1999, *Peeling the onion: An approach to comparing vessels glass assemblages*, «Journal of Roman Archaeology», 12, 72-100.

Cool H.E.M., Baxter M.J. 2005, *Cemeteries and significance tests*, «Journal of Roman Archaeology», 18, 397-403.

de Leeuw J., Mair P. 2009, *Simple and canonical Correspondence Analysis using the R package anacor*, «Journal of Statistical Software», 31, 1-18.

Djindjian F. 1985, *Seriation and toposeriation by Correspondence Analysis*, in A. Voorrips, S.H. Loving (eds.), *To Pattern the Past*, «PACT», 11, 119-135.

Drennan R.D. 2009, *Statistics for Archaeologists. A Commonsense Approach*, New York, Springer.

Everitt B.S., Landau S., Leese M., Stahl D. 2011, *Cluster Analysis*, 5th ed., Chichester, Wiley.

Gillis C. 1990, *Minoan Conical Cups. Form, Function and Significance*, Studies in Mediterranean Archaeology, 89, Goteborg.

Glynn D. in press, *Correspondence Analysis. Identifying for patterns of correlation*, in D. Glynn, J. Robinson (eds.), *Polysemy and Synonymy. Corpus Methods and Application in Cognitive Semantics*, Amsterdam, John Benjamins (http://www.academia.edu/attachments/27292432/download_file).

Goring-Morris A.N., Bergman C.A. 1987, *The Levantine Aurignacian with special reference to Ksar-Akil, Lebanon*, «Paléorient», 13, 142-147.

Greenacre M. 1988, *Clustering the rows and columns of a Contingency Table*, «Journal of Classification», 5, 39-51.

Greenacre M. 2007, *Correspondence Analysis in Practice*, Boca Raton-London-New York, Chapman&Hall/CRC.

Greenacre M. 2011, *The Contributions of Rare Objects in Correspondence Analysis*, Barcellona GSE Working Paper Series, Paper n. 571, 1-21.

Hair J.F., Black W.C., Babin B.J., Anderson R.E. 2009, *Multivariate Data Analysis*, 7th ed., Prentice Hall.

Healey J.F. 2013, *The Essentials of Statistics. A Tool for Social Research*, 3rd ed., Belmont, Wadsworth.

Husson F., Josse J., Pagès J. 2010, *Principal Component Methods-Hierarchical Clustering-Partitional Clustering: Why Would we Need to Choose for Visualizing Data?*, Technical Report-Agrocampus, Applied Mathematics Department, 1-17.

Husson F., Lê S., Pagès J. 2011, *Exploratory Multivariate Analysis by Example Using R*, Boca Raton-London-New York, CRC Press.

Ihaka R., Gentelman R. 1996, *R: A language for data analysis and graphics*, «Journal of Computational and Graphical Statistics», 5, 299-314.

Jackson D.A. 1993, *Stopping rules in Principal Components Analysis: A comparison of heuristical and statistical approaches*, «Ecology», 74, 2204-2214.

Kjeld Jensen C., Høilund Nielsen K. 1997, *Burial data and Correspondence Analysis*, in C. Kjeld Jensen, K. Hoilund Nielsen (eds.), *Burials & Society. The Chronological and Social Analysis of Archaeological Burial Data*, Aarhus, Aarhus University Press, 29-61.

Kuijt I., Goodale N. 2009, *Daily practice and the organization of space at the dawn of agriculture: A case study from the Near East*, «American Antiquity», 74, 403-422.

Lê S., Josse J., Husson F. 2008, *FactoMineR: An R package for multivariate analysis*, «Journal of Statistical Software», 25, 1-18.

Lorenzo-Seva U. 2011, *Horn's parallel analysis for selecting the number of dimensions in Correspondence Analysis*, «Methodology», 7, 96-102.

Madsen T. 1989, *Seriation and multivariate statistics*, in S. Rahtz, J. Richards (eds.), *Computer Applications and Quantitative Methods in Archaeology*, BAR International Series 548, Oxford, Archaeopress, 205-214.

Mameli L., Barceló J.A., Estevez J. 2002, *The statistics of archaeological deformation processes. An archaeozoological experiment*, in G. Burenhult, J. Arvidssen (eds.), *Archaeology in the Age of the Internet*, Oxford, Archaeopress, 1-17.

Milligan G.W., Cooper M.C. 1985, *An examination of procedures for determining the number of clusters in a data set*, «Psychometrica», 5, 159-179.

Morris J. 2008, *Associated bone groups; One archaeologist's rubbish is another's ritual deposition*, in O. Davis, K. Waddington, N. Sharples (eds.), *Changing Perspectives on the First Millennium BC*, Oxford, Oxbow, 83-98.

Nenadic O., Greenacre M. 2007, *Correspondence Analysis in R, with two- and three-dimensional graphics. The ca package*, «Journal of Statistical Software», 20, 1-13.

Pavùk P. 2010, *Pottery processing at Troy. Typology, stratigraphy and Correspondence Analysis. How do they work together*, in B. Horejs, R. Jung, P. Pavùk (eds.), *Analysing Pottery. Processing-Classification-Publication*, Gondova, Comenius University in Bratislava, 73-98.

Peeples M.A., Schachner G. 2012, *Refining Correspondence Analysis-based ceramic seriation of regional data sets*, «Journal of Archaeological Science», 38, 2818-2827.

Pitts M. 2005, *Pots and pits: Drinking and deposition in Late Iron Age south-east Britain*, «Oxford Journal of Archaeology», 24, 143-161.

Potter J.M. 2000, *Pots, parties, and politics: Communal feasting in the American Southwest*, «American Antiquity», 65, 471-492.

Preacher K.J., Zhang G., Kim C., Mels G. 2013, *Choosing the optimal number of factors in Exploratory Factor Analysis: A model selection perspective*, «Multivariate Behavioral Research», 48, 28-56.

Rakotomalala R. 2013, *Tanagra-Correspondence Analysis* (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Correspondence_Analysis.pdf).

Rowntree D. 2000, *Statistics without Tears: An Introduction for Non-Mathematicians*, London, Penguin Book.

Saporta G. 2006, *Probabilités, analyse des données et statistique*, 2^{e} ed., Paris, Editions Technip.

Shennan S. 1997, *Quantifying Archaeology*, Edinburgh, Edinburg University Press.

Smith K.Y., Neiman F.D. 2007, *Frequency seriation, Correspondence Analysis, and Woodland period ceramic assemblage variation in the Deep South*, «South Eastern Archaeology», 26, 47-72.

Taylor R. 1990, *Interpretation of the correlation coefficient: A basic review*, «Journal of Diagnostic Medical Sonography» 1, 35-39.

Van Pool T.L., Leonard R.D. 2011, *Quantitative Analysis in Archaeology*, New York, Wiley-Blackwell.

Wallin P. 2010, *In search of rituals and groups dynamics: Correspondence Analysis of Neolithic grave fields on the Islands of Gotland in the Baltic Sea*, «Documenta Praehistorica», 37, 65-75.

Weller S.C., Romney A.K. 1990, *Metric Scaling. Correspondence Analysis*, Sage University Paper 75, Newbury Park-London-New Delhi, Sage.

Wilson P., Cooper C. 2008, *Finding the magic number*, «The Psychologist», 21, 866-867.

Worthington R.L., Whittaker T.A. 2006, *Scale development research. A content analysis and recommendations for best practices*, «The Counselling Psychologist», 34, 806-838.

Yelland P.M. 2010, *An introduction to Correspondence Analysis*, «The Mathematica Journal», 12, 1-23.

Zwick W.R., Velicer W.F. 1986, *Comparison of five rules for determining the number of components to retain*, «Psychological Bulletin», 99, 432-442.

ABSTRACT

Over the years Correspondence Analysis has become a valuable tool for archaeologists because it enables them to explore patterns of associations in large contingency tables. While commercial statistical programs provide the facility to perform Correspondence Analysis, a number of packages are available for the free R statistical environment. Nonetheless, its command-line structure may be intimidating for users and prevent them from considering the technique. This article describes an R script, written by the author, which aims to free the R user from manually entering long pieces of code. By discussing two worked examples, it shows how the script can provide the user with a body of graphical and textual outputs relevant to the interpretation of data structure. It is hoped that the script will allow the user to concentrate more on the analysis results rather than the syntax of the R environment.