QÜESTIIÓ, Vol 18, 1 pp. 51-73, 1994

BUILDING A KNOWLEDGE BASE FOR CORRESPONDENCE ANALYSIS[1]

Mª CARMEN BRAVO LLATAS[†]

Universidad Complutense de Madrid

This paper introduces a statistical strategy for Correspondence Analysis. A formal description of the choices, actions and decisions taken during data analysis is built. Rules and heuristics have been obtained from the application of this technique to real case studies.

The strategy proposed checks suitability of certain types of data matrices for this analysis and also considers a guidance and interpretation of the application of this technique. Some algorithmic-like rules are presented and specific criteria are given for application.

This strategy has been implemented in a statistical knowledge-based system prototype using hypertext and rules.

Key words: Statistical Strategy, Correspondence Analysis, Knowledge-Based System, Knowledge Enhancement System, Hypertext.

[1] This paper has been partially supported by Spanish CICYT Project INF91-74.

[†] Mª Carmen Bravo Llatas. Centro de Proceso de Datos. Universidad Complutense. Avda. Paraninfo s/n. 28040 Madrid. Spain.

–Article presentat al Seventh International Conference on Multivariate Analysis, setembre 1992.

–Acceptat el gener de 1994.

1. INTRODUCTION

The increasing number of user friendly statistical packages together with the daily use of Statistics by non-experts may lead to its misuse. This is the main reason why statistical knowledge-based systems or statistical knowledge enhancement systems (Hand (1987, 1990)) are necessary. A first aim of these systems should be to help researchers in data analysis tasks providing them with guidance to choose a technique for analysis, to perform a successful analysis and to interpret the obtained results. Another goal of these systems should be the knowledge improvement in a statistical technique both by non-experts and experts in Statistics.

When dealing with the creation of a statistical knowledge base for a system, attention could be paid to two types of knowledge: the statistical theoretical background about a technique or a set of techniques and, the heuristics and rules that data analysts adopt in their work. In this paper, we concentrate on the latter and propose a statistical strategy for simple Correspondence Analysis. It also can be applied when more than two variables are present. The multiple Correspondence Analysis can be reduced to the two-way analysis partitioning the variable set into two groups, each of them composed of independent variables. In this case, a submatrix of the Burt matrix is analised, with different row and column categories, Lebart *et al.* (1984), Benzécri and Maïti (1990), Benzécri (1990).

A statistical knowledge enhancement system prototype has been developed combining these two types of knowledge. Rules and hypertext have been the tools employed to represent statistical knowledge. A deeper view of this system may be seen in Ferrán (1991), who organizes knowledge about Correspondence Analysis in order to implement it in a hypertext, and Ribes (1991), who describes the prototype. In Bravo (1991) a methodology for building a statistical strategy is proposed, and a part of the statistical strategy presented here is outlined.

For further information on these systems where hypertext is one tool employed to represent knowledge, see Hand (1987, 1990). Concerning strategies for Correspondence Analysis, Jambu (1991) gives some rules of action for application.

The rest of the paper is organized as follows: In Section 2, the statistical strategy concept is defined and a general view of Correspondence Analysis strategy is outlined. In Section 3, determining suitability of matrix substrategy is described, with emphasis on variables and categories. Two trees representing the main points of the strategy are also presented. Sections 4 to 6 show strategies for selection of the number of axes to be retained, interpretation based on one space and simultaneous interpretation. Some algorithms with specific criteria are presented and some ideas on how to implement them in the prototype are outlined. Section 7 presents some information about the implementation of the prototype together with a short example.

2. STATISTICAL STRATEGY

Hand (1986) defines a statistical strategy as a formal description of the choices, actions and decisiones taken during data analysis. It is not merely a sequence of actions but a recipe to choose an action as a result of an external event.

There are two well-known types of statistical strategies depending on their main purpose. One is related to the selection of some statistical techniques suitable for analysing a specific data set and the other provides a guidance to the analysis and interpretation of results, once a statistical technique has been chosen. For the development of our strategy, Correspondence Analysis is assumed to be the technique chosen to analyse data.

As many authors agree (Lubinsky and Pregibon (1988), Thisted (1986), Olford and Peters (1986), Pregibon (1986)), a statistical strategy may be represented by a set of trees where every node is a different representation of data. Furthermore, data analysis consists in navigating through this set where any movement from one node to another is done by means of any transformation, such that: a test, a complicated procedure, a researcher's preference or even data themselves.

Hand (1986), Huber (1986) and Pregibon (1986) consider that data analysis does not merely consist in data cleaning followed by analysing but in these four stages:

(a) **Objective formulation.** In the application context researcher's objectives are formulated and suitable techniques are considered. Dependent and independent variables are determined as well. In our case, there is no need to select a technique because Correspondence Analysis has been chosen for application. Even though, we will determine if our main aim is basically descriptive or scores on axes will be used in a subsequent analysis; select active and illustrative variables; combine categories to form illustrative ones; or make a decision for missing value categories.

(b) **Formalization.** In this stage, the translation of objectives to statistical terms is carried out.

(c) **Numerical processing**. It consists in modeling test, data cleaning and selection of data transformations. Undoubtedly this is the most valuable part a knowledge enhancement system should contain. At this stage, we would give criteria or algorithms to select axes, intepret them or determine well represented categories. Some data cleaning is also included such as, for example, if only a few categories explain the maximum inertia axis we should reconsider the matrix to be analysed.

(d) **Interpretation of results**. In subject-matter context, we will interpret retained axes and proximities between points projected onto axes.

Although these stages are followed in a sequential way it is not a rigid structure. Iterative or recursive cycles may occur in numerical processing. Besides, from numerical processing and from the interpretation it is possible to go back to the first stage to look for another objective. Similarly, if disparity between objectives and results is encountered a reformulation of an old objective may be done.

Note also that interaction between the system and the user may take place in the four stages. Whenever it is possible, the user's preferences should be taken into account to choose the way analysis should proceed.

As in most data analysis tasks, I believe that Correspondence Analysis strategy has hierarchical structure that can be represented by a tree. It is subdivided in two main interrelated parts connected to the two types of statistical strategy: determining suitability of the matrix to analyse and, guiding and interpreting the analysis. The interrelation between these two hierarchical parts can be explained by displacements from a node of one branch to the other branch. The first strategy is composed of two substrategies: one forms the data matrix to be analysed and the other determines when the analysis is appropriate for this data set. The second strategy is composed of four substrategies: selection of the number of axes to be retained, interpretation based on one space, simultaneous interpretation and interpretation of interesting and illustrative categories. These substrategies are also broken down into some substrategies and so on.

3. DETERMINING SUITABILITY OF MATRIX

3.1. Determining data matrix

Since Correspondence Analysis is an exploratory data analysis technique revealing interrelations between two spaces of categories, we shall usually use it to analyse large data matrices of positive numbers. This is motivated by the difficulty of drawing significant information just from viewing these matrices.

This substrategy, presented in figures 1 and 2, deals with the configuration of data matrix for analysis, with regard to variables and categories. It also implies some data cleaning and selects illustrative variables or categories to be projected onto the axes.

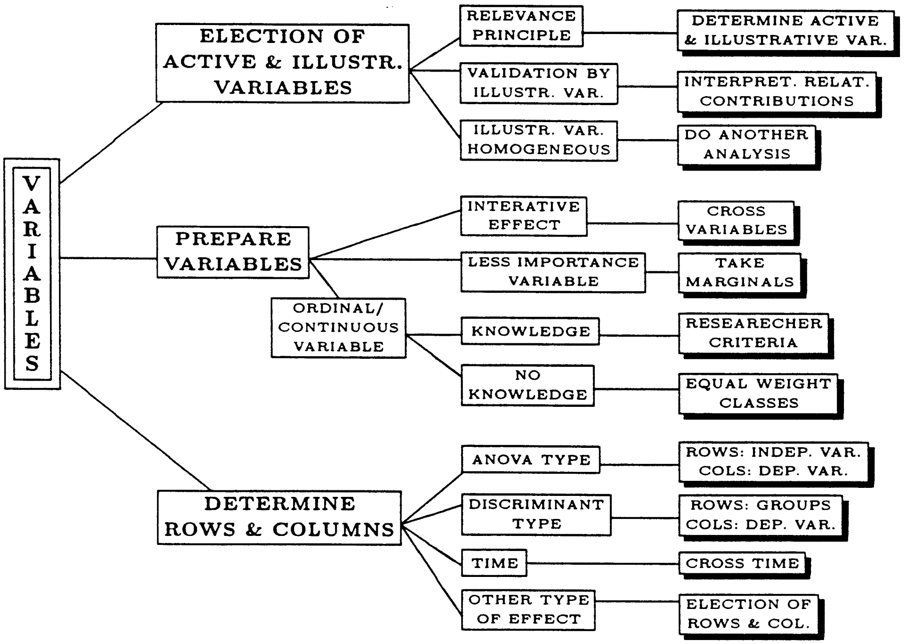


Figure 1.
Statistical strategy for variables.

3.1.1. Variables

Globally, the substrategy related to variables deals with the selection of active and illustrative variables, as well as the preparation of these variables and the determination of row and column spaces of the matrix to be analysed.

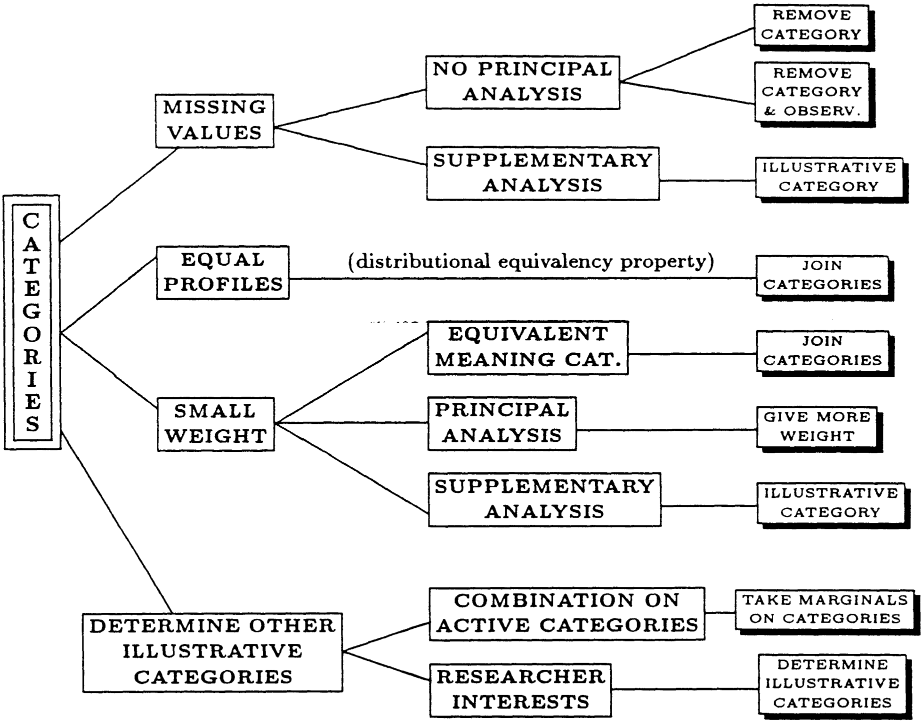


Figure 2.
Statistical strategy for categories.

Taking into account the application context, the steps below should be followed to specify that matrix:

(a) Select active and illustrative variables following the Relevance Principle to preserve homogeneity. In Correspondence Analysis, we require not only measure homogeneity of variables, but also their semantic homogeneity. The Relevance Principle should be applied to keep for principal analysis those variables interrelated in a specific point of view. Not being illustrative variables in the analysis, they could be used as a validation test for principal analysis, Lebart *et al.* (1984) and Benzécri *et al.* (1990).

(b) If there are continuous variables, categorize them following the researcher's knowledge or any other criteria as, for example, subdividing in similar weight classes, Tekaïa *et al.* (1990).

(c) If there are ordinal variables, consider their categorization if the number of possible values is greater than a fixed number, i.e., four or five. It is convenient to follow the reserarcher's interests and not to unbalance the weight of classes. Even, extreme value categories could be joined when it is judged reasonable.

(d) If more than two active variables are present:

- ▶ determine sets of variables with interactive effects on the others. Afterwards, variables in each of these sets should be crossed to create a new active one, which replaces those. Commonly, a temporal variable will be crossed with those of interest in evolution, Benzécri *et al.* (1990), Alawieh (1990), Maravalle (1990).
- ▶ determine variables of less important to take marginals on them.
- ▶ determine row and column spaces of the matrix to be analysed, which is built extracting the corresponding submatrices of the Burt matrix.

(e) Even when Correspondence Analysis is an exploratory technique, sometimes it is useful to apply it to data matrices traditionally analysed by other confirmatory techniques, such as, for example, analysis of variance, nonparametric multiple comparison tests or discriminant analysis. Although these techniques did not reveal significant effects, Correspondence Analysis may show important interrelations among variables, which are easy to check with the data matrix, Benzécri *et al.* (1990), Maïti (1989), Tekaïa *et al.* (1990).

3.1.2. Categories

This substrategy considers the treatment given to special types of categories such as, missing value, similar profile and small weight and determines illustrative categories. It is subdivided in these parts:

(a) For missing value categories, the strategies that could be followed are: define them as illustrative categories, keep them in principal analysis or remove them from principal analysis. In case of elimination of categories from principal analysis we also could consider the elimination of experimental units with missing values on them.

(b) In each space, similar profile categories will be joined, as a consequence of distributional equivalence property of $\chi-$squared distance, Lebart *et al.* (1982).

(c) If there are small weight categories, reconsider their presence in principal analysis. For each of them, three strategies may be followed: join the category to another of similar meaning, define the category as an illustrative one or give more weight to it.

A category could be considered of small weight when its weight is lower than $1/(\alpha \times p)$, p being the number of categories corresponding to the variable and α a real number greater than 1, e.g., 3.

(d) Define other illustrative categories taking marginals on combinations of active categories or considering the researcher's interests.

3.2. Determining suitability of analysis

The aim of Correspondence Analysis is the distribution of total inertia in a few axes. Although it could be applied in the independent case, it is more valuable when rows and columns of the matrix are not independent. In this case, the analysis tries to explain interrelations among row and column points. Note, however, that if the distribution of total inertia on axes were homogeneous, Correspondence Analysis would not be the best suitable descriptive technique for these data matrices, Lebart *et al.* (1984).

4. SELECTION ON THE NUMBER OF AXES TO BE RETAINED

The selection of the number of axes to be retained depends on our main purpose, if it is merely descriptive or if in subsequent analysis the original variables are going to be replaced with scores on the axes.

The proposed strategy, which is not based on a statistical test, tries to get a compromise among well-known criteria. These criteria are: a) to retain only a few axes, b) to keep as much of total inertia as possible, c) to avoid that the inertia of any retained axis being similar to the inertia of any other non-retained axis, d) to retain those axes whose proportion of inertia is greater than the 'inertia mean'.

Algorithm below combines specific versions of those aspects:

```
Algorithm 1.

Step 0. Initialization:
        $\Lambda (= 0)$ : inertia explained by retained axes.
        $k (= 0)$ : number of retained axes.
        $\epsilon (= 0.25|0.20)$ : coefficient to determine a big step between two eigenvalues.
        $m, s$ : degrees of freedom of rows and columns.
        $k'$ : maximum number of axes to be retained.
        $\lambda_k$ : $k^{\text{th}}$ eigenvalue or inertia explained by $k^{\text{th}}$ axis.
        $\Sigma\lambda_l$ : total inertia minus one.

Step 1. $k = k + 1$
        $\Lambda = \Lambda + \lambda_k$
Step 2. If $\left( \left( \frac{\Lambda}{\Sigma\lambda_l} \geq 0.80 \text{ and } |\lambda_{k+1} - \lambda_k| > \epsilon \times \lambda_k \right) \text{ OR } (k \geq \min(m, s)) \right)$
        then go to Step 3.
        else go to Step 1.

Step 3. If $k > k'$ then        Reinitialization: $k = 0$
                                                  $\Lambda = 0$
            3.1. $k = k + 1$
            If $\lambda_k > \frac{\Sigma\lambda_l}{\max(m, s)}$        then $\Lambda = \Lambda + \lambda_k$
                                                                        Go to 3.1.
            else $k = k - 1$
                 Go to Step 4.
        else go to Step 4.

Step 4. Results:
        $k$ : number of axes to be retained.
        $\frac{Lambda}{\Sigma\lambda_l}$ : proportion of inertia explained by retained axes.
        End.
```

In this algorithm, all computations are supposed to start at the second eigenvalue because the first one is always one and it doesn't make sense.

5. ONE SPACE INTERPRETATION

This strategy deals with the fact of interpretation based on one space. Usually, for each retained axis the row or column spaces will be chosen to interpret it. Afterwards, the projections on the axis of the other space points will be interpreted, taking into account that badly represented categories over the retained axes set should lead us to reconsider their presence in the analysis.

5.1. Interpretation of an axis

The algorithm introduced for the intepretation of an axis provides criteria to select the categories that explain mostly the axis according to the category absolute contributions. It also determines the relative importance of these categories in this explanation and locates the category projections on both sides of the axis. Note that we should also include those categories whose absolute contributions are similar to one of those of the explicative categories. Lebart *et al.* (1982), define the absolute contribution of a point to an axis as the proportion of total inertia of the axis explained by the point.

The algorithm also considers the case when very few categories explain the maximum inertia axis. In this case, the matrix to be analysed should be reconsidered, according to the following substrategies for those categories: join them to another category of similar meaning, remove them from principal analysis or give them less weight.

Once chosen one of the two spaces, the algorithm proposed to interpret an axis is as follows:

Algorithm 2.

Step 0. *Initialization:*

LIMIT $(= 1/n)$: limit to retain an explicative category.
n : number of categories in the space.
$k(= 1)$: number of categories explaining the axis.
$\theta(= 0.5)$: coefficient to detect similarities between absolute contributions.
$\alpha(= 0.20)$: coefficient to obtain N.
N : maximum number of categories that explain the first axis prior to reconsidering the matrix to be analysed.
$N = \text{INT}[n \times \alpha]$ if $n < 20$; $N = 4$ if $n \geq 20$.

```
    $\text{CABS}_i(i = 1, \ldots, n)$ : absolute contribution of $i^{\text{th}}$ point to the axis.
    $\text{PROJ}_i(i = 1, \ldots, n)$ : projection of $i^{\text{th}}$ point onto the axis.

Step 1. Order categories decreasingly according to their absolute contributions:
        $\text{CABS}_1 > \cdots > \text{CABS}_n$
        Verify that $\text{CABS}_1 > \text{LIMIT}$
        $\text{SIGN} = \text{SIGN}(\text{PROJ}_1)$

Step 2. $k = k + 1$
        If $\text{CABS}_k < \text{LIMIT}$ then go to Step 5.

Step 3. $I = 1$
        3.1. If $(\text{CABS}_1 - \text{CABS}_k) < \theta \times \text{LIMIT} \times I$ then the relative importance
               of the two points in the construction of the axis is $I$-similar.
               Go to Step 4.

             else $I = I + 1$
                  Go to 3.1.

Step 4. If $\text{SIGN}(\text{PROJ}_k) = \text{SIGN}$ then the $k^{\text{th}}$ point is projected on the same
        semiaxis of the point of greatest importance.

        else $k^{\text{th}}$ and the point of greatest importance are in opposite semiaxes.

        Go to Step 2.

Step 5. $k = k - 1$
        If $(\text{CABS}_{k+1} - \text{CABS}_k) < (\theta/2) \times \text{LIMIT}$
        then readjust the number of categories, $k$, due to similarity between
             absolute contributions.

Step 6. If it is the first axis and $k < N$ then reconsider the configuration of the
        matrix following the strategy given above the algorithm.
        Go to End.

        else  Results:
              the first $k$ categories explain the axis.
              their relative importance is given in Step 3.
              their relative position on the axis is given in Step 4.
              End.
```

The criteria specified in the algorithm may be substituted as, for example, the limit for retaining categories could be LIMIT $= \mu/(n-1)$, being $\mu = 0.75$ or any other constant smaller than one.

Practically speaking, in the prototype built, the set of categories that explain an axis is divided only into three groups of importance. The interval between the limit to accept explicative categories and the greatest absolute contribution among categories is divided into three equal size subintervals.

5.2. Representation of categories on an axis

The algorithm proposed for intepreting categories on an axis provides criteria to select well represented categories on the axis according to the category relative contributions. It also positions the projections of well represented categories on both sides of the axis and considers the selection of categories whose relative contributions are similar to one of those of the well represented categories. Beforehand, the presence of badly represented categories in the retained axes set should lead us to reconsider their presence in the analysis. According to Lebart *et al.* (1982) relative contribution of an axis to a point is the squared cosine of the angle formed by the point with the axis.

Note that retaining a category to interpret an axis (based on absolute contribution), position it on a semiaxis and not to represent this category onto it because of a low relative contribution, may be confusing. It may be due to the fact that this category would participate in the explanation of some other axes, Lebart *et al.* (1984).

Once interpreted an axis by the categories of one space, the algorithm for interpretation of the other space projections onto it is given by:

Algorithm 3.

Step 0. *Initialization:*

LIMIT$(= 0.40)$: minimum value for relative contributions to select a well represented category.

CAT: the other space category of greatest absolute contribution.

SIGN: sign of the projection of CAT.

n: number of categories in the space.

$k(= 1)$: number of well represented categories on the axis.

$\text{CREL}_i(i = 1, \ldots, n)$: relative contribution of axis to i^{th} point.

$\text{PROJ}_i(i = 1, \ldots, n)$: projection of i^{th} point onto the axis.

$\alpha(= 0.05)$: coefficient to detect similarities between relative contributions.

```
Step 1. Order categories decreasingly according to their relative contributions:
        $\mathrm{CREL}_1 > \cdots > \mathrm{CREL}_n$.
        Verify that $\mathrm{CREL}_1 > \mathrm{LIMIT}$.
Step 2. $k = k + 1$
        If $\mathrm{CREL}_k < \mathrm{LIMIT}$ then go to Step 5.
Step 3. If $\mathrm{SIGN}(\mathrm{PROJ}_k) = \mathrm{SIGN}$ then the $k^{\mathrm{th}}$ point and CAT are on the same
        semiaxis.
        else the $k^{\mathrm{th}}$ point and CAT are in opposite semiaxes.
Step 4. $I = 1$
        4.1. If $(1 - \mathrm{CREL}_k) < \alpha \times I$ then the level of good representation of $k^{\mathrm{th}}$
             point is $I$.
                 Go to Step 2.

             else $I = I + 1$
                 Go to 4.1.
Step 5. $k = k - 1$
        If $(\mathrm{CREL}_{k+1} - \mathrm{CREL}_k) < \alpha/2$ then readjust the number of categories,
        $k$, diminishing or augmenting it.
Step 6. Results:
        The first $k$ categories can be explained by the axis.
        The way they are correlated with the axis and other space categories
        is given by relative contributions and by Step 3.
        The levels of good representation are given in Step 4.
        End.
```

All the specified criteria in this algorithm may be modified. For example, the limit in order to select well represented categories, could be related to the global quality of representation of categories in the retained axes set, provided that these qualities are acceptable. This limit could be $\mathrm{LIMIT} = \mathrm{Q}/\mathrm{NRA}$ or $\mathrm{LIMIT} = 1/\mathrm{NRA}$, where Q is the global quality of representation of the category and NRA is the number of retained axes.

In our prototype, the set of well represented categories is divided into three groups whose relative contributions lie in [0.40, 0.60), [0.60, 0.80) and [0.80, 1], respectively. Considering that the limits of these groups depend on the dimensionality of the data, they could be changed depending on the number of inertia axes in each problem.

6. SIMULTANEOUS REPRESENTATION AND INTERESTING AND ILLUSTRATIVE CATEGORIES INTERPRETATION

Simultaneous interpretation is done by explaining proximities between projected points of both spaces onto axes. It should be taken into account their quality of representation and weights, as well as Euclidean distances between points and from the origin, which is simultaneously the center of gravity of row and column points.

For well represented categories on an axis, the proximity of a row (column) point to a column (row) point is encountered when the component of this column (row) in the row (column) profile has a higher value than the same component in the average row (column) profile. The row (column) profile components with lower values than the corresponding components in the average row (column) profile are the column (row) categories which are projected on different semiaxis that the row (column) category. Similarly, two row (column) points are close to each other when their profiles are similar and are located on different semiaxes when they have very different profiles. In fact, the proximity between two row (column) points is explained by higher values in their profiles than the average row (column) profile of the components corresponding to the nearest column (row) points, and lower values than the average row (column) profile of components corresponding to the other semiaxis column (row) points.

The above interpretations will be more accurate when points are farther from the origin, better represented, and have more weight, Lebart *et al.* (1984). It would be interesting to investigate ways of combining these criteria in order to obtain a measure of the relative importance of these interpretations.

For categories that are illustrative or of interest for the researcher, the axes in which they are well represented have to be selected. Proximities to other well represented categories on those axes may be interpreted as the above paragraphs suggest. Section 5 also may be applied to intepret projections onto axes.

In the prototype built, simultaneous representation of well represented points on axes is shown, and some guidelines are given to the user for interpretation.

7. IMPLEMENTATION OF THE PROTOTYPE. AN EXAMPLE

A prototype for PC MS-DOS has been built. Both the statistical strategy and the theoretical background of the technique have been included. There is a continuous interaction between the user and the prototype, allowing them to participate in matrix building and analysis and to learn how the analysis is performed and interpreted by the prototype. The user's knowledge about the technique also can be improved using the hypertext and explaining commands.

Tools employed in the prototype are: a) KES^R II[2] (Knowledge Engineering System II, Release 2.2), an expert system tool, b) HIPER1, the hypertext, written in C (Mosich *et al.*, 1988), c) ANCO, a FORTRAN program that performs the analysis. KES^R II is composed of a rule-based knowledge base and a backward chaining inference engine.

KES^R II helps us to build a menu-driven interface between the user and the prototype. It displays two main menu screens: One to determine the matrix for analysis and another to perform the analysis; both of them include several submenu screens and questions. It is possible to ask for explanation of technical terms about those questions or reasons leading to them by the EXPLAIN and WHY commands.

Both menu screens allow for linking to HIPER1, which contains the theoretical background of the technique in textual cards. HIPER1 allows the navigation through its knowledge base in a non-sequential way, letting the user to establish interrelations among statistical terms according to their needs. Graphical aids would make more understandable some concepts and strategies, and could be included in later versions.

During matrix construction, the prototype tries to obtain some metainformation available of data. It suggests the application of Relevance Principle to select active and illustrative variables; inquires into dependencies among variables (anova, discriminant, time effect, others) and interactive effect of some variables on the others; and asks for: the user's interests in categorization of ordinal and continuous variables, how to deal with non-complete cases, missing value and small weight categories, and the determination of some illustrative categories as for example marginals and others of user's interest.

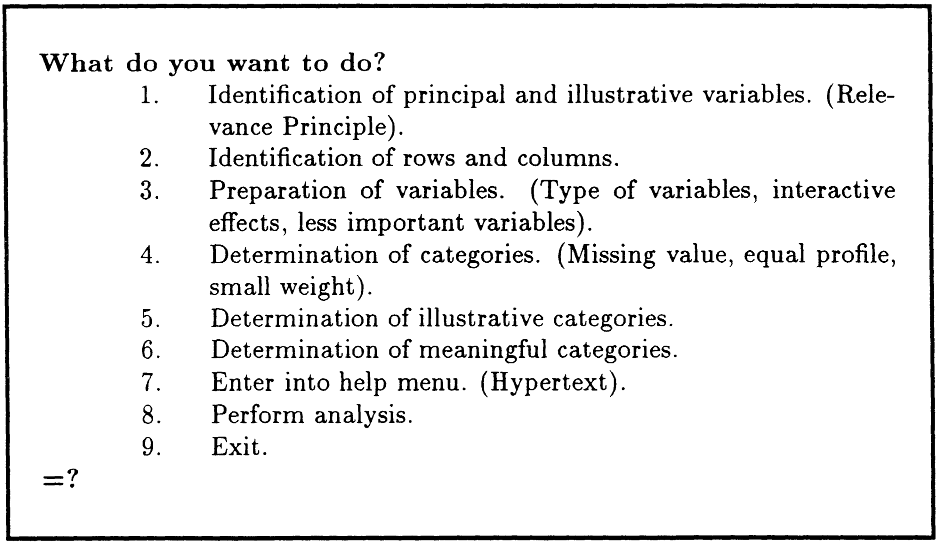
Once built the data matrix, the analysis is performed by ANCO, which creates some communication files with KES^R II, which applies the strategy.

Having proved the success of the conjunction of hypertext, rules and programming techniques in the construction of the prototype, we suggest the development of a more powerful system. We would suggest also the use of a hypertext

[2] KES^R II (c) (1985) Software Architecture & Engineering, Inc., distributed and supported by Sperry Corp.

as the interface between the user and the system and the incorporation of tools that would allow the system to get as metainformation of data as possible.

The following short example illustrates the use of the prototype. It concerns a study of the influence of body position in some respiratory measures, taking into account sex, age, body mass index (bmi), smoking and exercise. The initial menu screen is shown in figure 3.



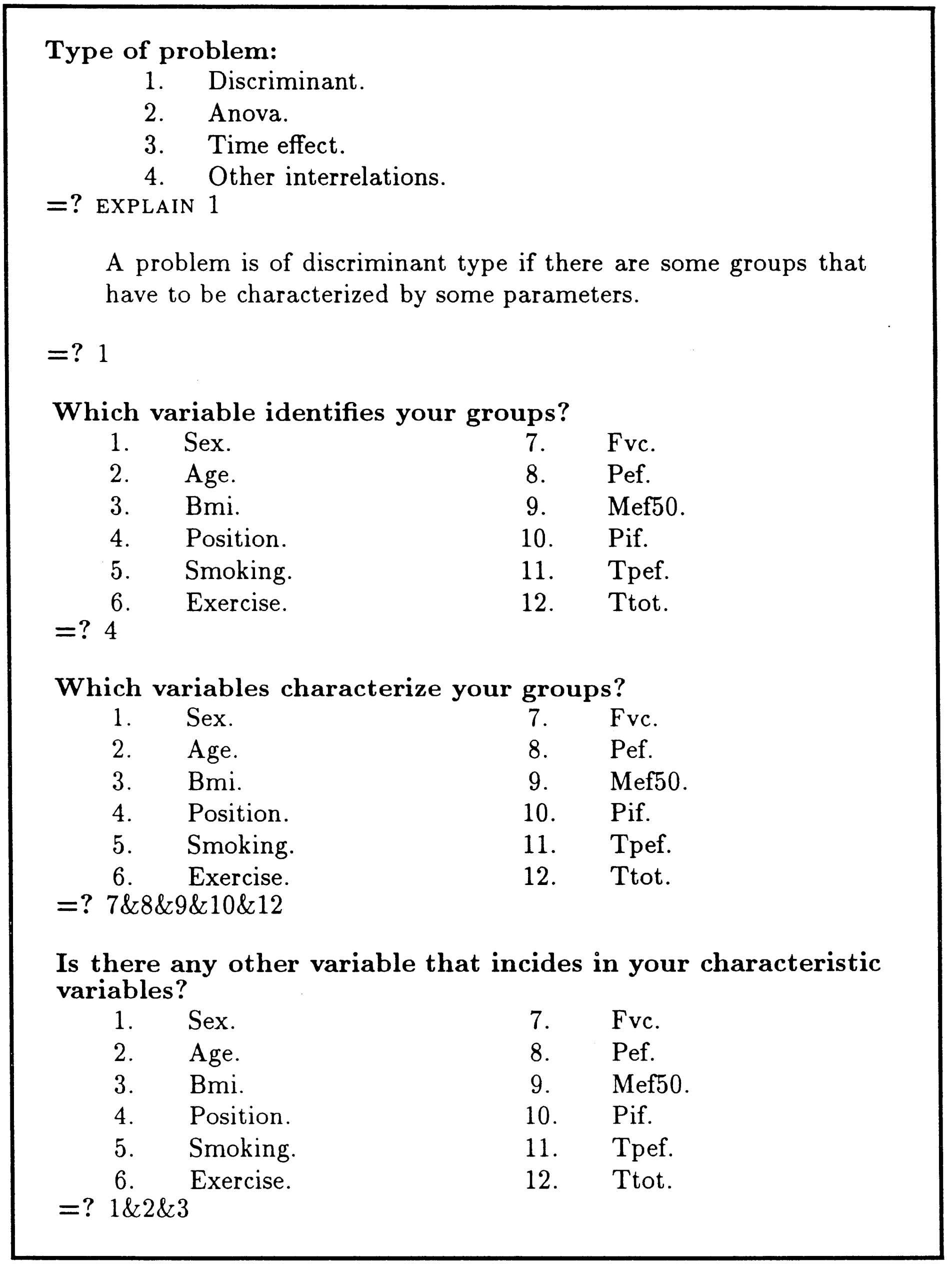
What do you want to do?

1. Identification of principal and illustrative variables. (Relevance Principle).
2. Identification of rows and columns.
3. Preparation of variables. (Type of variables, interactive effects, less important variables).
4. Determination of categories. (Missing value, equal profile, small weight).
5. Determination of illustrative categories.
6. Determination of meaningful categories.
7. Enter into help menu. (Hypertext).
8. Perform analysis.
9. Exit.

=?

Figure 3.
Menu for determining the matrix for analysis.

The matrix for analysis will be determined by clicking points displayed in this menu. Through point 1, the prototype asks for selection of active variables among the observed ones, according to the Relevance Principle: *sex, age, bmi, position, smoking, exercise and some respiratory parameters. Smoking and exercise* are considered unimportant. Point 2, see figure 4, finds a discriminant variable, *position*; variables characterizing the groups, *respiratory parameters*; and other variables influencing on these, *sex, age and bmi.* Point 3 helps to detect an interactive effect of *sex and age on respiratory parameters*; determines the categorization of continuous variables in three classes: *age* according to the user's interests, *bmi and respiratory parameters* in equivalent weight categories, defining different *bmi* limits for each *sex*; and variables of less importance, *smoking, exercise and some respiratory parameters.* Point 4 considers the presence of missing value, small weight and equal profile categories.

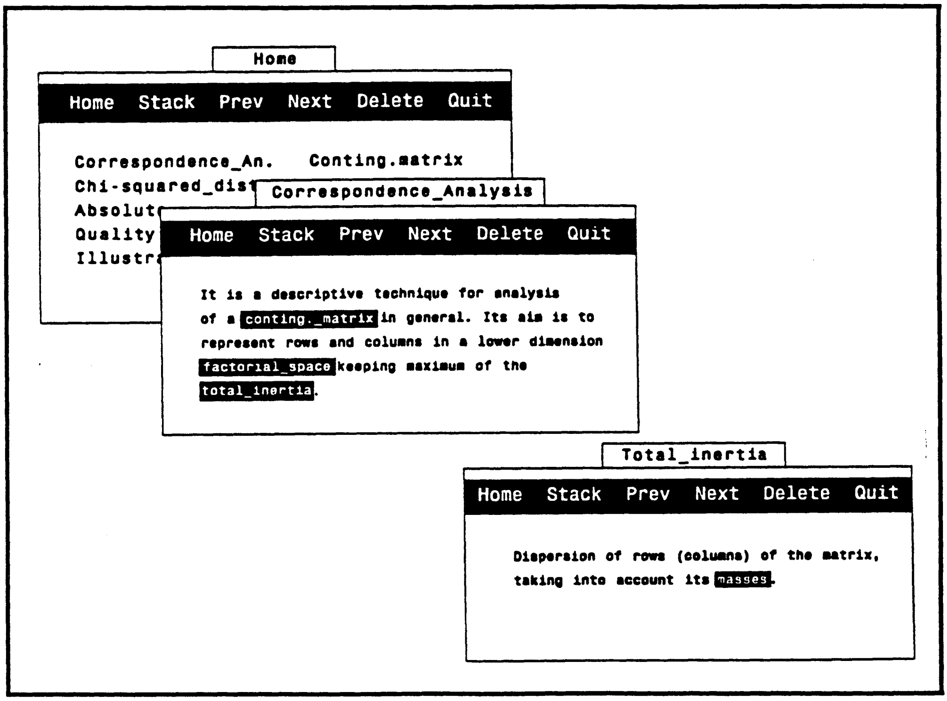


```
Type of problem:
          1.   Discriminant.
          2.   Anova.
          3.   Time effect.
          4.   Other interrelations.
=? EXPLAIN 1

     A problem is of discriminant type if there are some groups that
     have to be characterized by some parameters.

=? 1

Which variable identifies your groups?
      1.    Sex.                    7.    Fvc.
      2.    Age.                    8.    Pef.
      3.    Bmi.                    9.    Mef50.
      4.    Position.              10.    Pif.
      5.    Smoking.               11.    Tpef.
      6.    Exercise.              12.    Ttot.
=? 4

Which variables characterize your groups?
      1.    Sex.                    7.    Fvc.
      2.    Age.                    8.    Pef.
      3.    Bmi.                    9.    Mef50.
      4.    Position.              10.    Pif.
      5.    Smoking.               11.    Tpef.
      6.    Exercise.              12.    Ttot.
=? 7&8&9&10&12

Is there any other variable that incides in your characteristic
variables?
      1.    Sex.                    7.    Fvc.
      2.    Age.                    8.    Pef.
      3.    Bmi.                    9.    Mef50.
      4.    Position.              10.    Pif.
      5.    Smoking.               11.    Tpef.
      6.    Exercise.              12.    Ttot.
=? 1&2&3
```

Figure 4.
Identification of rows and columns of the matrix.

Point 5 defines some illustrative categories, such as marginals on *sex and age and other categories.* Finally, point 6 determines meaningful categories to be interpreted. This menu helps to determine the matrix for analysis. Here, it is the submatrix extracted from the Burt matrix with rows the categories of *position*, *bmi* and the combination of *sex and age*, and columns, the categories of the *respiratory parameters* of interest in principal analysis.

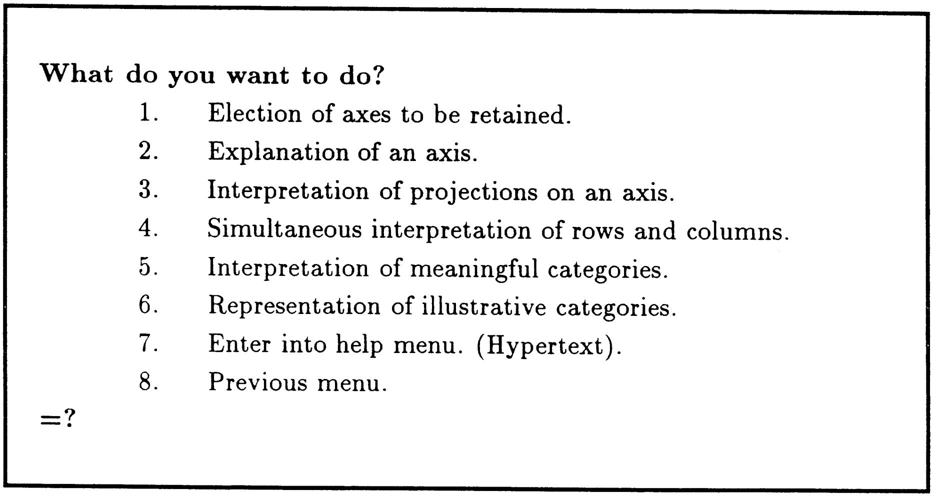
Figure 5 displays the hypertext, which may be called from the two main menu screens. Once built the matrix for analysis, point 8 of the initial menu leads to the second manin menu screen shown in figure 6. Point 1 of this second menu shows the histogram of proportion of inertia explained by the axes and gives some explanation about the number of axes that are retained in this step. Other points are self explanatory enough. Figures 7 and 8 display part of the output of the session, but not the dialogue between the user and the prototype. Output in figure 7 is related to points 2 and 3 of this second menu, while figure 8 is related to point 4.



It is a descriptive technique for analysis of a conting._matrix in general. Its aim is to represent rows and columns in a lower dimension factorial_space keeping maximum of the total_inertia.

Dispersion of rows (columns) of the matrix, taking into account its masses.

Figure 5.
The hypertext.



What do you want to do?

1. Election of axes to be retained.
2. Explanation of an axis.
3. Interpretation of projections on an axis.
4. Simultaneous interpretation of rows and columns.
5. Interpretation of meaningful categories.
6. Representation of illustrative categories.
7. Enter into help menu. (Hypertext).
8. Previous menu.

=?

Figure 6.
Menu for performing the analysis.

8. CONCLUDING REMARKS

The main goal of developing statistical strategies is to build statistical knowledge enhancement systems that not only lead to a better use of data analysis techniques but also to the improvement of user's knowledge about them.

The developed strategy and the organized knowledge about Correspondence Analysis have been implemented in a knowledge enhancement system prototype. Next step should be the transformation of this prototype into a modular system to which new modules related to other techniques could be added subsequently. This modular system would represent a valuable tool to analyse data correctly and to learn or improve users' knowledge in data analysis techniques.

According to our experience, these systems not only help researchers in other fields to perform analysis but also benefit Statistics and data analysts. Statistics is protected against misapplication of techniques to data and data analysts may improve their own knowledge about certain techniques.

(**Point 2.** Determined an axis and row points to explain it)

First axis is explained by the categories:
$mc^-, mc^+, pos1, pos2$, and $pos3$
Below, groups of categories by their importance in determining the axis (absolute contributions):
Maximum importance categories: $mc^+, pos1$, and $pos3$
Medium importance categories: mc^-, and $pos2$
Moreover, the categories: $mc^+, pos1$, and $pos2$
are opposed in this axis to the categories: mc^-, and $pos3$

Ready for command:

(**Point 3.**)

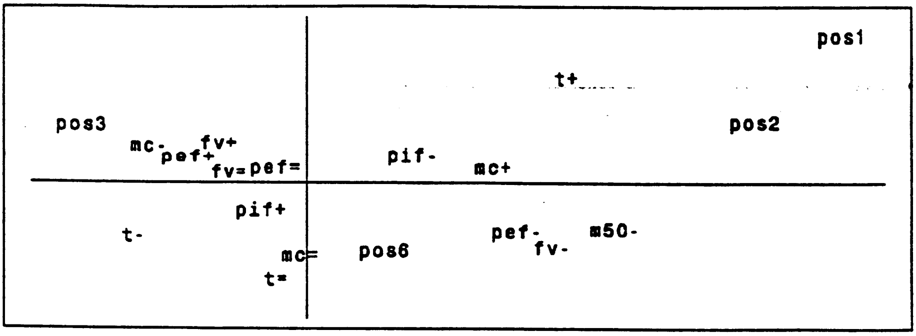
Being explained this axis by row categories, column categories will be explained.
The categories well represented onto first axis are:
$fv^-, fv^=, fv^+, pef^-, pef^=, pef^+, m50^-, m50^=, m50^+, pif, pif^+, t^-$, and t^+
Below, groups of categories by levels of good representation (relative contributions):
Rather well represented categories: $fv^-, pef^-, m50^-, m50^+$, and t^-
Quite well represented categories: $fv^=, fv^+, pef^+, pif^+$, and t^+
Fairly well represented categories: $pef^=, m50^=$, and pif^-
Moreover, the categories: fv^-, pef^-, and t^+
are projected onto positive semiaxis being correlated with the categories: $mc^+, pos1$, and $pos2$
and the categories: $fv^=, fv^+, pef^=, pef^+, m50^-, m50^=, m50^+, pif^-, pif^+$, and t^-
are projected onto negative semiaxis being correlated with the categories: mc^-, and $pos3$

Ready for command:

Figure 7.
Parts of the output displayed by the prototype during analysis.

(Point 4.)

Categories well represented onto first (horizontal) and second (vertical) axes:



For simultaneous interpretation of well projected categories in each axis take into account:

- One row (column) point is near one column (row) point because the corresponding column (row) component in the row (column) profile is higher than this component in the average row (column) profile; the column (row) points on the other semiaxis have lower values in their row (column) profile components than the corresponding components in the average row (column) profile.
- The proximitity between two row (column) points is derived from higher values than the average row (column) profile of the components corresponding to the nearest column (row) points, and lower values than the average row (column) profile of the components corresponding to the other semiaxis column (row) points.
- The more weighty and farther the points are from the origin the more accurate these interpretations are.
- It should be taken into account quality of representation of points.

Figure 8.
Part of the output displayed by the prototype in point 4 of second menu.

ACKNOWLEDGEMENTS

The author wishes to thank Professor José Miguel García-Santesmases, Professor Magdalena Ferrán and technician Belén Ribes without whose help and skills this work would not be done.

REFERENCES

[1] **Alawieh, A.** (1990). "Répartition en France par Régions de la Collecte pour L'Ensemble des Céréales". *Les Cahiers de L'Analyse des Données*, **XV**, **3**, 367–370.

[2] **Benzécri, J.P.** (1990). "Analyse des Données Biologiques et Patholologie Clinique". *Les Cahiers de L'Analyse des Données*, **XV**, **3**, 285–304.

[3] **Benzécri, J.P.** and **Maïti, G.D.** (1990). "Etude des Réactions de 2000 Sujets à une Thérapeutique". *Les Cahiers de L'Analyse des Données*, **XV**, **4**, 463–474.

[4] **Benzécri, J.P., Maïti, G.D.** and **Kremer, C.** (1990). "Sevrage d'un Traitement Hypnotique ou Anxiolytique et Thérapeutique Substitutive". *Les Cahiers de L'Analyse des Données*, **XV**, **4**, 447–462.

[5] **Bravo, M.C.** (1991). "Estrategia Estadística en el Sistema de Mejora del Conocimiento MACABE". *Trabajo de Investigación. Departamento de Estadística e Investigación Operativa.* Universidad Complutense de Madrid.

[6] **Ferrán, M.** (1991). "Generación de Bases de Conocimientos para Sistemas de Mejora del Conocimiento. Aplicación al Sistema MACABE". *Trabajo de Investigación. Departamento de Estadística e Investigación Operativa.* Universidad Complutense de Madrid.

[7] **Hand, D.J.** (1986). "Patterns in Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 355–387.

[8] **Hand, D.J.** (1987). "A Statistical Kwnoledge Enhancement System". *The Journal of the Royal Statistical Society*, **Series A (General), Vol. 150, Part 4**, 334–345.

[9] **Hand, D.J.** (1990). "Emergent Themes in Statistical Expert Systems". *Knowledge, Data and Computer-Assisted Decisions*, **Nato Asi Series, Vol. F 61**, Springer-Verlag, 279–288.

[10] **Huber, P.J.** (1986). "Environments for Supporting Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 285–294.

[11] **Jambu, M.** (1991). *Exploratory and Multivariate Data Analysis.* Academic Press.

[12] **Lebart, L., Morineau, A.** and **Warwick, K.M.** (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices.* John Wiley & Sons.

[13] **Lebart, L., Morineau, A.** and **Fénelon, J.P.** (1982). *Traitement des Données Statistiques.* Dunod.

[14] **Lubinsky, D.** and **Pregibon, D.** (1988). "Data Analysis as Search". *Journal of Econometrics,* **38**, 247–268.

[15] **Maïti, G.D.** (1989). "Etude Comparative de L'Efficacité de Deux Médicaments Antalgiques et de Leur Association". *Les Cahiers de L'Analyse des Données*, **XIV**, **2**, 157–162.

[16] **Maravalle, M.** (1990). "Géographie Politique de L'Italie D'Après les Votes à Quatre Scrutins Nationaux de 1983 à 1989". *Les Cahiers de L'Analyse des Données*, **XV**, **2**, 191–208.

[17] **Mosich, D., Shammas, N.** and **Flamig, B.** (1988). *Advanced TURBO C^R Programmer's Guide.* Willey.

[18] **Oldford, R.W.** and **Peters, S.C.** (1986). "Implementation and Study of Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 335–353.

[19] **Pregibon, D.** (1986). "A DIY Guide to Statistical Strategy". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 389–399.

[20] **Ribes, B.** (1991). "Un sistema de Representación del Conocimiento Estadístico. Sistema MACABE". *Trabajo de Investigación. Departamento de Estadística e Investigación Operativa.* Universidad Complutense de Madrid.

[21] **Tekaïa, F., Sansonetti, Ph.** and **Claverie, J.M.** (1990). "Estimation du Estade de L'Infection par le VIH Chez les Sujets Séro-positifs". *Les Cahiers de L'Analyse des Données*, **XV**, **3**, 261–278.

[22] **Thisted, R.A.** (1986). "Representing Statistical Knowledge for Expert Data Analysis Systems". Gale, W.A., Ed., *Artificial Intelligence and Statistics*, Addison-Wesley, Reading, Mass., 267–283.