Statistica & Applicazioni - Vol. VIII, n. 1, 2010, pp. 27-56

AN INTEGRATED APPROACH TO REGRESSION ANALYSIS USING CORRESPONDENCE ANALYSIS AND CLUSTER ANALYSIS

Jules J. de Tibeiro*
Luigi D'Ambra**

SUMMARY

Problems involving dependent pairs of random variables usually involve two aspects: tests of independence or estimation of measures of association. In order to find out which way best explains the data, this paper addresses Regression Analysis applied to Correspondence Analysis (CA). It also uses Agglomerative Hierarchical Clustering as a method to accompany Multiple Correspondence Analysis (MCA). A well known data set is analyzed.

Keywords: *Complete Disjunctive Table, Burt Matrix, Regression Table, Multiple Correspondence Analysis, Agglomerative Hierarchical Clustering.*

1. INTRODUCTION

There have been considerable developments of statistical methods that analyze, in a single framework, the asymmetrical relationships of more than one subset of quantitative variables (criterion and explanatory), as well as qualitative variables.

The mathematical analysis of data is often divided into two antagonistic schools. On the one hand, there are the adepts of *traditional statistics* (tests of hypotheses, analysis of variance, regression analysis, general linear model, etc.). On the other hand, there is *data analysis* using scaling techniques that include two principal areas: *Correspondence Analysis* (CA) and *Hierarchical Clustering* (HC).

This, in fact, leads to *Classification*, i.e., the identification of *homogeneous* and *distinct subgroups* in data where one focuses on an aspect (sequence of aggregations) of the computational algorithm used. Here *Hierarchical Clustering* is intended as used as a companion method to CA: *Agglomerative Hierarchical Clustering* (in French: ''*Classification Ascendante Hiérarchique*'' (CAH)).

Note that *Classification* is an extremely broad subject area which often requires *Discriminant Analysis*. In other words, we first want to know in which respects the given groups of subjects are different. Then we apply this knowledge to classify new subjects into groups.

Typically, the data tables to be analyzed are made of several measurements, col-

* Secteur des Sciences - Université de Moncton - Campus de Shippagan - SHIPPAGAN, NB, Canada E8S 1P6 (e-mail: jules.de.tibeiro@umcs.ca).

** Dipartimento di Scienze Biologiche - Università di Napoli Federico II - via Mezzocannone, 8 - 80134, NAPOLI (e-mail: dambra@unina.it).

lected from a set of units (e.g., subjects). In general, the units are rows (observations, objects or individuals) and the variables are columns (questions). The main goals of *Regression Analysis* can be summarized as data description (to investigate or refute a relationship among variables), interpretation using a fitted model to obtain an interpolation or calibration curve/surface) and finally, inference.

A common procedure when using applied regression analysis, is to select a subset of the available predictor variables and to estimate regression coefficients of the subset by Ordinary Least Squares (OLS) estimates. In other words, *a subset of the regression coefficients is reduced to zero*, while the rest are estimated by least squares and are therefore regarded as not at all reduced. This often leads to alterations of the initial model such as transformations of the data or further regression techniques.

The purpose of this paper arises from our questions related to the practice of correlation and regression analysis. The bonds existing between two variables are usually so complex that it is not judicious to express them with only one number. Moreover, in our view, the possibility of expressing one or more *response variables* within a group of *explanatory variables* could be possible using CA on a contingency table and thus crossing the categories of some of these variables, as well as others.

Benzécri (1992) and Cazes (1977) presented more than adequate methods for dealing with solutions offered by *polynomial regression* as well as showing how to progress step by step in order to protect data quality. They suggested a judicious prior compression of data by substituting numerous primary variables with a reduced number of *coordinates* constituting as large a base (all observations) and as stable a base (very few variables) as possible.

Many practical studies follow the following scheme: a set of individuals (observations) I is described by a set of variables J which one can subdivide into a set of *explanatory variables* X and a set of *response variables* Y. The problem is to *find* and *explain* relationships (causal or not) between the variables of X and those of Y.

In general, if Y is *reduced to only one variable*, y (which is the case in the present study), several traditional methods of prediction are applicable, according to the type of the variable y and to the following types of variables of X:

- *regression analysis* is possible if y is *quantitative* and the variables of X are *quantitative* and also *categorical* considering only the *dummy variables* (indicator variables) that can be treated as *quantitative variables*;
- *analysis of variance*, if y is *quantitative*, and if all variables of X are *qualitative* (we also get this case, by dividing variables into categories, if certain variables of X are *quantitative*);
- *discriminant analysis*, if y is *qualitative*, and the variables of X are *quantitative* and also *categorical* considering only the *dummy variables* (indicator variables) that can be treated as *quantitative variables*;
- *barycentric discrimination*, if y and all variables of X are *qualitative* (one can, in the same way, again get this case if certain variables of X are quantitative).

If Y is *not reduced to only one variable*, one can use the factorial methods (*Reduced Rank Regression, Principal Component Regression, Principal Component*

Analysis, Correspondence Analysis, etc. according to different cases). For more details, see D'Ambra, Amenta and Gallo (2005).

When there is a small number of observations for which *both explicative variables and response variables are known*, keeping a test sample (that will not be used to calculate the regression coefficients, but rather to choose the number of explanatory (predictor) variables to be preserved) leads to a doubtful reduction of the basis for the calculation. See Cazes (1975) and Brenot (1977).

As suggested by Cazes (1977), we study in this paper a regression problem between a *response variable* and a set of *categorical predictor variables*. For more details, see Cazes (1997), and de Tibeiro (1997).

When there is a large number of predictors, one can obtain a model with too many parameters and consequently a model which models the error. To avoid such an over-parametrization, the *Partial Least Square* (PLS) *regression* may be introduced. As the PLS components depend on the connection between the *response variables* and *the predictors*, we cannot calculate the variances of the regression coefficients with a simple formula. For more details, see Tenenhaus (1998) and Cazes (1997).

We will proceed as follows in this paper. We present, in Section 2, a short mathematical background of CA and its ''natural'' connection with hierarchical clustering. In Section 3, we propose a methodology of regression *with* CA for the case study presented in Section 4. Results and concluding remarks are given respectively in Sections 5 and 6.

2. PREREQUISITE NOTIONS OF CORRESPONDENCE ANALYSIS

2.1 *Algebra of Two-Way Correspondence Analysis*

Correspondence Analysis (CA) is an exploratory computational method for the study of associations between variables. CA can be used to analyze several types of multivariate data. All involved some categorical variables. Much like *Principal Component Analysis* (PCA), it displays a low-dimensional projection of the data, e.g., into a plane.

The objective of (two-way) CA is to portray data geometrically as a set of row and column points in, say, two-dimensional space for ease of visualization. Let rows, I, and columns, J, be collected into the $I \times J$ data matrix $\mathbf{N}$ (with elements n_{ij}) representing a contingency table of two *categorical variables* with positive row and column sums (almost always $\mathbf{N}$ consisting of nonnegative numbers, but there are some exceptions, such as the one described at the end of Chapter 23 in Greenacre (2007)).

Let n_{i+} and n_{+j} denote the sum of the i-th row and j-th column, respectively, and $n_{++} = \sum_i \sum_j n_{ij} = \mathbf{1}^T\mathbf{N}\mathbf{1}$ denote the grand total of $\mathbf{N}$. The notation $\mathbf{1}$ is used here for a vector of ones of length that is appropriate to its use; hence the first $\mathbf{1}$ is $I \times 1$ and the second is $J \times 1$ to match the row and column lengths of $\mathbf{N}$. The mass of the

i-th row is defined as $r_i = n_{i+}/n_{++}$ and likewise the mass of the j-th column is $c_j = n_{+j}/n_{++}$. We note respectively $\mathbf{r}$ and $\mathbf{c}$, the vector of *row masses* and the vector of *column masses*.

The matrix $\mathbf{N}$ is first converted to the so-called *correspondence matrix* of relative frequencies $\mathbf{P}$ by dividing $\mathbf{N}$ by its grand total n_{++} as $\mathbf{P} = \mathbf{N}/n_{++}$ with entries $p_{ij} = n_{ij}/n_{++}$. The following notation is used respectively *for row and column masses*: $r_i = \sum_{j=1}^{J} p_{ij}$ i.e., $\mathbf{r} = \mathbf{P1}$ and $c_j = \sum_{i=1}^{I} p_{ij}$ i.e., $\mathbf{c} = \mathbf{P}^T\mathbf{1}$. Let $\mathbf{D}_r = diag(\mathbf{r})$ and $\mathbf{D}_c = diag(\mathbf{c})$ denote respectively the *diagonal matrices of row and column masses*.

Note that all subsequent definitions and results are given in terms of these relative quantities $\mathbf{P} = \{p_{ij}\}$, $\mathbf{r} = \{r_i\}$ and $\mathbf{c} = \{c_j\}$, whose elements add up to 1 in each case. Multiply these by n_{++} to recover the elements of the original matrix $\mathbf{N}$: $n_{++}p_{ij} = n_{ij}$, $n_{++}r_i = i$-th row sum of $\mathbf{N}$, $n_{++}c_j = j$-th column sum of $\mathbf{N}$.

We define the *row profiles* as the *rows* of the original table $\mathbf{N}$ divided by respective row totals, equivalently $\mathbf{D}_r^{-1}\mathbf{P}$. Similarly, we define the *column profiles* as the columns of the original table $\mathbf{N}$ divided by respective column totals, equivalently $\mathbf{P}\mathbf{D}_c^{-1}$. The assumption of *independence* is

$$p_{ij} = r_i c_j, i = 1, \ldots, I, j = 1, \ldots, J$$

All of CA is based on the *computational algorithm* to obtain *coordinates* of the *row* and *column profiles* with respect to *principal axes*. The first step of this algorithm is the computation of the so-called *matrix* $\mathbf{S}$ *of standardized residuals*:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-1/2} \tag{1}$$

with elements $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$. The second step of this algorithm is the computation of the *Singular Value Decomposition* (SVD) of

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T \tag{2}$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, $\mathbf{D}_\alpha$ is the diagonal matrix of (positive) *singular values* in descending order: $\alpha_1 \geq \alpha_2 \geq \ldots$ and columns of matrices $\mathbf{U}$ and $\mathbf{V}$ are *left* and *right singular vectors*, respectively.

The following steps in CA are to define the configuration's *row and column coordinates*, $\mathbf{F}$ and $\mathbf{G}$ (after dropping the dimensionality subscript), as follows:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha = \Phi\mathbf{D}_\alpha, \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha = \Gamma\mathbf{D}_\alpha \tag{3}$$

where the *standard coordinates* Φ *of rows* and the *standard coordinates* Γ *of columns* are respectively defined as

$$\Phi = \mathbf{D}_r^{-1/2}\mathbf{U}, \Gamma = \mathbf{D}_c^{-1/2}\mathbf{V} \tag{4}$$

The two sets of coordinates, $\mathbf{F}$ and $\mathbf{G}$, represent the final set of outputs of interest to the researcher. They usually are plotted together in a two (or more) dimensional space. According to the SVD of $\mathbf{S}$, the *squares of singular values* (α_k^2) or *principal inertias* (λ_k) of $\mathbf{S}$ also decompose total inertia as

$$\lambda_k = \alpha_k^2, k = 1, 2, \ldots, K \text{ where } K = \min\{I - 1, J - 1\}$$

To understand the link between CA and the *biplot*, we need to introduce a mathematical formula which expresses the original data n_{ij} in terms of the row and column masses and its coordinates. From the relations (1), (2) and (4) of the basic computational algorithm, the data in $\mathbf{P}$ can be written as

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^{K} \sqrt{\lambda_k} \phi_{ik} \gamma_{jk}\right)$$

This is known as *the bilinear CA model*, also called the *reconstitution formula*, which can be written in matrix notation as

$$\mathbf{P} = \mathbf{D}_r(\mathbf{1}\mathbf{1}^T + \Phi \mathbf{D}_\lambda^{1/2} \Gamma^T)\mathbf{D}_c$$

Because of the simple relations (3) between the principal and standard coordinates, this bilinear model can be written in several alternative ways. For more details, see Benzécri (1992) and Greenacre (2007).

The *left* and *right singular vectors* are related linearly, for example by multiplying the SVD on the right by $\mathbf{V}$: $\mathbf{SV} = \mathbf{U}\mathbf{D}_\alpha$. Expressing such relations in terms of the *principal* and *standard coordinates* gives the following variations of the same theme, called *transition equations* that govern the *asymmetric maps* for the equivalent scalar versions:

Principal as a function of standard (barycentric relationships):

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{P}\Gamma, \mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}^T\Phi \tag{5}$$

Principal as a function of principal:

$$\mathbf{F} = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{G}\mathbf{D}_\lambda^{-1/2}, \mathbf{G} = \mathbf{D}_c^{-1}\mathbf{P}^T\mathbf{F}\mathbf{D}_\lambda^{-1/2} \tag{6}$$

The equations (6) express the *profile points* as weighted averages of the vertex points, where the weights are the *profile elements*. They govern the *asymmetric maps*. They show that the two sets of principal coordinates, which govern the *symmetric map*, are also related by a *barycentric* (weighted average) relationship, but with scale factors (the inverse square roots of the principal inertias) that are different on each axis.

The total inertia of the data matrix is the sum of squares of the matrix $\mathbf{S}$ in (1) and also the sum of squares of the singular values, i.e., the sum of the eigenvalues:

$$\begin{aligned} \textit{Inertia} = \textit{trace}(\mathbf{S}\mathbf{S}^T) = \textit{trace}(\mathbf{S}^T\mathbf{S}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}(p_{ij} - r_i c_j)^2/(r_i c_j) \\ &= \sum_{k=1}^{K}\alpha_k^2 = \sum_{k=1}^{K}\lambda_k \end{aligned}$$

In other words, the *part of variance* of the cloud accounted for by axis k is equal

to the *eigenvalue* λ_k. The proportion of variance $\lambda_k \Big/ \sum_{k=1}^{K} \lambda_k$ accounted for by axis k is a descriptive index of the importance of axis k.

To first approximation, CA can be understood as an extension of PCA where the variance in PCA is replaced by an inertia proportional to the chi-square distance of the table from independence. CA decomposes this measure of departure from independence along axes that are orthogonal according to a chi-square inner product. If we are comparing two categorical variables, the simplest possible model is that of independence in which case the counts in the table would obey approximately the margin products identity.

As is well known, CA provides useful *aids to interpretation*, among which the *eigenvalues* (λ_k, $k = 1, \ldots, K$ associated with the percentages τ_k, $k = 1, \ldots, K$) and *principal coordinates* (basic results of CA), obtained by computer software. One deduces: the *relative contributions of points to axes* and the *quality of representation* of each object on each object factorial axis. The quality of representation is defined as the square of the cosine of the angle that the object forms with the object axis.

For more details including the *transition equations* defined in (5) and (6), see Benzécri (1973,1992), Lebart *et al.* (1997), Greenacre (1984), Murtagh (2005) and Le Roux and Rouanet (2004).

2.2 *Overview of Multiple Correspondence Analysis*

Multiple Correspondence Analysis (MCA) is defined as an extension of CA to more than $Q = 2$ variables, which allows one to analyze the pattern of relationships of several *categorical dependent variables*. Suppose the original matrix of categorical data is $N \times Q$, i.e., N cases and Q variables.

The first form of MCA converts the cases-by-variables data to an *indicator matrix* $\mathbf{S}$ where the categorical data have been recorded as *dummy variables*. If the q-th variable has J_q categories, this indicator matrix will have $J = \sum_q J_q$ columns.

Then the indicator version of MCA is the application of the basic CA algorithm defined above in Section 1 to the matrix $\mathbf{S}$, resulting in coordinates for the N cases and the J categories.

The second form of MCA calculates the $J \times J$ table obtained as $\mathbf{B} = \mathbf{S}^T \times \mathbf{S}$ of all two-way cross-tabulations of the Q variables and is called the *Burt table* (or Burt matrix). Then the Burt version of MCA is the application of the same basic CA algorithm to the *symmetric matrix* $\mathbf{B}$, resulting in coordinates for the J categories.

The *standard coordinates* of the categories are identical in both versions of MCA, and the principal inertias in the Burt version are the squares of those in the indicator version. Moreover, the eigenvalues obtained from CA of the Burt table give, in general, a better approximation of the inertia, explained by the factors, than the eigenvalues of $\mathbf{S}$. For more details, see Murtagh (2005) and Greenacre (2007).

2.3 *Agglomerative Hierarchical Clustering*

The aim of a *Cluster Analysis* is to derive a partition, or a sequence of partitions, of a set of objects, based on their similarities (equivalently, their distances) to one another, so that objects clustered into the same group (or class) are similar, or close, to one another, while those of different groups are dissimilar, or far apart. In *Agglomerative Hierachical Clustering*, the I objects are regarded initially as I clusters of one object each, and the analysis proceeds sequentially to lump together clusters into larger clusters, until all the objects form a single cluster.

Let's remember that in a proper sub-space (plane 1×2, space $1 \times 2 \times 3$, etc.) stemming from the analysis of a *correspondence table* **N**, the set I can appear divided in known classes before the analysis, but for which the composition is not explicitly noted in table **N**.

Most of the classification algorithms, and particularly the *agglomerative algorithms*, are locally robust in the sense that the lower parts of the produced dendrograms are largely independent of possible outliers. For all these reasons, particularly in the case of large data sets (which is not the case in our study), it is highly advisable to complement CA with a classification performed on the whole space, or at least in the high-dimensional space spanned by all *the significant axes*. See Benzécri (1992).

In other words, it is often more efficient to perform a classification using a *limited number of factors* issued from CA. We note that a technique of hierarchical clustering such as the reciprocal neighbour algorithm (McQuitty, 1966), and particularly the chain search algorithm (Benzécri, 1997a,b) can be performed without storing the array of distances in the central memory. See also Juan (1982) and Murtagh (2002).

The most significant categories or variables of variables characterizing each cluster are automatically selected and sorted, therefore producing a computer-aided description of the classes, and hence, of the whole multidimensional space. See Lebart, Mourineau and Warwick (1984). See also Murtagh (2005), Le Roux and Rouanet (2004), Lebart (1997) and Jambu (1983).

It appears useful to combine CA with another inductive method intended to provide not a spatial representation but an *automatic classification*. As a result of the chosen distance, the method works well with CA: it is possible to produce from the total variance of the cloud $N(I)$ (representing the set I to classify) a double decomposition following the nodes of the clustering and the axes retained in CA. This allows us to combine interpretation of both clusters and factors. For more details, see Benzécri (1992), Lebart *et al.* (1997), Greenacre (1984) and Murtagh (2005).

In order to reduce the complexity of the study, we propose to perform a \ *preliminary* agglomerative hierarchical clustering as a prior condensation of the data. This clustering is carried out on an *indicator matrix* crossing all the observations, while splitting up explanatory variables and the response variable. It uses an algorithm which is founded on the property of *reducibility*. We consider this algorithm that may be used to build up a hierarchy of classes, and then concentrate on the criterion of inertia, which closely fits the χ^2-distance used in CA. For more details, see Lebart (1994).

3. REGRESSION ANALYSIS WITH MCA

Our methodology is at the heart of MCA, studying the regression problem between a *response variable* and a set of *categorical predictor variables*. We suppose that all the variables $x_1, x_2, \ldots, x_i, \ldots, x_p$, $y_1, y_2, \ldots, y_j, \ldots, y_q$, have been divided into classes, and we designate by K_{x_i} (*resp.* K_{y_j}) the set of categories of the variable x_i $(1 \leq i \leq p)$ [*resp.* y_j $(1 \leq j \leq q)$] and by K_X (*resp.* K_Y) the unconnected union of K_{x_i} (*resp.* K_{y_j}), i.e., the set of all the explanatory categories (resp. to explain):

$$K_X = \cup\{K_{x_i} | i = 1, \ldots, p\}; K_Y = \cup\{K_{y_j} | j = 1, \ldots, q\}$$

If E designates the set of n observations, then we consider the *complete disjunctive table (indicator matrix)* S_{EK_X} that we note simply S, associated with variables x_i of which the general term $S(e,k)$ is defined by

$$S(e,k) = \begin{cases} 1 \text{ if e has adopted the modality } k \text{ of } x_i \\ 0 \text{ if not} \end{cases}; \forall e \in E, \forall k \in K_{x_i} \subset K_X$$

We have the same notation for T_{EK_Y}, (or simply T), the *complete disjunctive table* associated with variables y_j. We designate $T(e,k)$ $(e \in E, k \in K_Y)$, as the general term of T. Regression analysis *with* CA involves carrying out the following steps:

Step 1: After dividing into slices of variables x_i and y_j, we construct the table $C = T^t S$ (associated with the *complete disjunctive table* $\mathbf{t}_{E(YJ)}$), which gathers together the set of qp contingency tables crossing every variable x_i $(1 \leq i \leq p)$ with every variable y_j $(1 \leq j \leq q)$.

Step 2: We carry out CA of data table C. We designate by $(\varphi_\alpha^{K_X}, \varphi_\alpha^{K_Y})$ the α^{th} couple of *factor coordinates* associated with variance 1 (derived from this analysis) and by λ_α the corresponding eigenvalue.

Step 3: We add the table S to supplement C, i.e., we project on the *r first factorial axes* (coordinates) found in Step 2 the profiles of the rows e in the table S. Let $F_\alpha(e)$ be the *factor coordinates* of the row's profile $e \in E$ on the factorial axis α.

Taking into account that $\sum\{S(e,k) | k \in K_X\} = p$, we obtain

$$F_\alpha(e) = \frac{1}{p} \sum_{k \in K_X} \varphi_\alpha^k S(e,k) \tag{7}$$

Step 4: We carry out the *usual regression} of each y_j (before dividing them into classes) with the factor coordinates} F_α.*

Let

$$y_j^*(e) = \sum_{\alpha=1,r} g_j^\alpha F_\alpha(e)$$

be the approximate value of y_j for the individual e. According to (7), we have

$$y_j^*(e) = \sum_{k \in K_X} b_j^k S(e,k)$$

with

$$b_j^k = \frac{1}{p} \sum_{\alpha=1,r} g_j^\alpha \varphi_\alpha^k$$

If $i(e)$ indicates the slice of x_i in which falls e

$$y_j^*(e) = \sum_{i=1,p} b_j^{i(e)}$$

This quite simply corresponds to a formula used in *analysis of variance*, insofar as the approximate value $y_j^*(e)$ of the variable y_j is obtained by adding the regression coefficients associated with the categories through the observation of e.

More precisely, this formula represents the *reconstitution* of y_j with a sum of terms associated with the modalities from the individual e, as is the case in analysis of variance.

For example, if y_{ijk} is the output of the fragment i with the type of ground j and the type of fertilizer k, we can write as $E(y_{ijk}) = a_j + b_k$, the sum of the terms associated with the modalities of this piece, this also remains true for the reconstitution y_{ijk}^* of y_{ijk} starting from the estimates of a_j and b_k.

Moreover, it is well known that if one carries out CA of the *complete disjunctive table* S_{EK_X} (denoted earlier simply as S) *by keeping all the factors (coordinates)*, the result of the regression is *identical* to the result of the *analysis of variance*. For more details about properties of regression after MCA, see Cazes (1977, 1997).

In classical linear regression, a set of n measurements (observations) on a dependent (predictand or response) variable y and on p independent (predictor) variables $x_1, \ldots, x_p$ is given. We consider the following centered data vectors of R^n: $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. Let $\mathbf{X}_{(n,p)} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ the data set associated with the predictor variables.

We denote

$$\mathbf{U_{XX}} = \mathbf{X}^T\mathbf{X} \text{ and } \mathbf{U_{Xy}} = \mathbf{X}^T\mathbf{y} \tag{8}$$

As $\mathbf{y}$ is an $n \times 1$ vector of the observations of the response variables, $\mathbf{X}$ is an $n \times p$ full rank matrix of levels of the regressor variables, β is a $p \times 1$ vector of the regression coefficients. The goal of regression analysis could be accomplished using ordinary multiple regression, which provides just one solution, often based on the *least squares criterion*. Thus, the least-squares estimator of β is

$$\hat{\beta} = (\mathbf{U_{XX}})^{-1}\mathbf{U_{Xy}} = (\mathbf{U_{XX}})^{-1}\mathbf{X}^T\mathbf{y}$$

Note that the $(\mathbf{U_{XX}})^{-1}$ matrix will always exist if the regressors are linearly independent.

PLS regression combines features from *Principal Component Analysis* (PCA) and multiple regression. Its goal is to predict or analyze a set of dependent variables $\mathbf{y}$ from a set of independent variables or predictors $\mathbf{X}$, and to describe their common structure. This prediction is achieved by extracting from the predictors a set of orthogonal factors called *latent variables* which have the best predictive power.

Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using stepwise methods). Another one, called *Principal Component Regression* (PCR), is to perform a PCA of the $\mathbf{X}$ matrix and then use the principal components (i.e., eigenvectors) of $\mathbf{X}$ as regressors on $\mathbf{y}$. Technically in PCA, $\mathbf{X}$ is decomposed, using its SVD as $\mathbf{X} = \mathbf{S}\Delta\mathbf{V}^T$ with $\mathbf{S}^T\mathbf{S} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ (matrices left and right singular vectors), and Δ being a diagonal matrix with the *singular values* as diagonal elements.

According to the previous relations from (8), we propose to carry out the regression on PLS components as follows. Specifically, the goal is to obtain a first pair of vectors:

$$\tau = \mathbf{X}[diag(\mathbf{U_{XX}})]^{-1}\mathbf{U_{Xy}} = \mathbf{Xc}, \text{ and } \mathbf{u} = \mathbf{yv}$$

where $\tau^T\mathbf{u}$ is maximal and $\mathbf{c} = [diag(\mathbf{U_{XX}})]^{-1}\mathbf{U_{Xy}} = [diag(\mathbf{U_{XX}})]^{-1}\mathbf{X}^T\mathbf{y}$ and

$$Var(\mathbf{c}) = [diag(\mathbf{U_{XX}})]^{-1}\mathbf{U_{XX}}[diag(\mathbf{U_{XX}})]^{-1}\sigma^2$$

When the first latent vector is found, it is subtracted from both $\mathbf{X}$ and $\mathbf{y}$ and the procedure is repeated until $\mathbf{X}$ becomes a null matrix. For more details, see the PLS regression algorithm (Tenenhaus, 1998).

Being satisfied with only one response variable, y, to simplify, if y is normally distributed, it is not the same for the PLS components, and thus, for the regression coefficients. The explanation of $\mathbf{y}$ by the PLS components τ may be done by

$$\hat{\mathbf{y}}_{PLS} = [(\tau^T\mathbf{y})/(\tau^T\tau)] \times \tau = \mathbf{d}\tau$$

where

$$\mathbf{d} = [(\tau^T\mathbf{y})/(\tau^T\tau)]$$

If t was independent of $\mathbf{y}$, one would have

$$Var(\mathbf{d}) = [\tau^T Var(\mathbf{y})\tau]/[\tau^T\mathbf{y}]^2 = \delta^2/[\tau^T\mathbf{y}]$$

But as τ depends on $\mathbf{y}$, the expression $\mathbf{d}$ is non-linear in $\mathbf{y}$, which ''intervenes'' in the numerator and the denominator of $\mathbf{d}$. That being the case, *one cannot calculate the variance of* $\mathbf{d}$. As a consequence, we can not determine a prediction interval unless we apply *bootstrap techniques*, which are used more and more when one does not have a simple formula to compute standard deviations.

The interest of the PLS approach is to force the connection between $\mathbf{y}$ and the predictors, but, it is a ''biased technique''. For example, when we want to explain the (quantitative) variables, based on the connection between the variables y_j $(j = 1, 2, \ldots, q)$ and the variables x_i $(i = 1, 2, \ldots, p)$ on which we will carry out the regression of each y_j. The goal of PLS regression is to provide new variables $\varphi_1, \varphi_2, \ldots, \varphi_r$ $(r \leq p)$ linear combinations of the not correlated explanatory variables.

The results obtained with PLS regression are ''biased'' because we use this approach to explain the y_j components which are already function of the connections

between the response variables y_j and explanatory variables x_i. For more details, see Cazes (1997).

4. THE DATA

We illustrate regression analysis, combined with MCA, with an example from a random sample of Gasoline Mileage Performance for automobiles during 1975 where the variables have a well-defined scale of measurement. Appendix Table B.3 in Montgomery *et al.* (2001,p. 570) presents gasoline mileage performance data on 32 automobiles, along with 1 response variable (*Miles/gallon)* and 11 taxonomic (explanatory) variables (Displacement, Horsepower, Torque, Compression ratio, Rear axle ratio, Carburetor, No. of transmission speeds, Overall length, Width, Weight, Type of transmission).

There are missing values in two of the observations (Trans AM and Star), so we will confine our analysis to only the 30 vehicles for which complete samples are available. We will retain 10 explanatory variables in a new dataset henceforth named Table 1.

We propose in the following rows the specification of the names (codes) assigned to the variables as they are found in the figures. We retained the first four letters to code each car: for example, Apollo (apol), Omega (omeg), Nova (nova), . . ., Corvette (corv). The cars Trans AM and Star were excluded because of missing values.

TABLE 1. - *Description and Coding of Variables*

Variable	Description	Coding
y	*Miles/gallon*	*MIL1, MIL2, MIL3, MIL4, MIL5*
x_1	*Displacement (cubic in.)*	*dis1, dis2, dis3, dis4*
x_2	*Horsepower (ft-lb)*	*hor1, hor2, hor3, hor4, hor5*
x_3	*Torque (ft-lb)*	*tor1, tor2, tor3, tor4, tor5*
x_4	*Carburetor (barrels)*	*car1, car2*
x_5	*No. of transmission speeds*	*nts1, nts2*
x_6	*Overall length (in.)*	*ovl1, ovl2, ovl3, ovl4, ovl5*
x_7	*Width (in.)*	*wid1, wid2, wid3, wid4*
x_8	*Weight (lb)*	*wei1, wei2, wei3, wei4, wei5*
x_9	*Type of transmission*	*typ*1: *automatic transmission (A)*
		*typ*2: *manual transmission (B)*

Source: Motor Trend (1975).

The aim of the study is to evaluate whether an automatic or manual vehicle has an effect on gasoline mileage performance. People who are driving manual vehicles will claim that they have tremendous gasoline mileage, saving money on gas, and concede that a full tank of gas can take them a longer distance than automatic vehicles.

However, let us remember that the objectives of multiple regression are somewhat different: one is trying to find an optimal (or near-optimal) ''classification'' structure, while the other seeks to develop a prediction equation. This study illustrates the use of CA to identify distinct profiles characterizing automatic or manual vehicle and the effect on gasoline mileage performance.

As the rows and columns in Table 1 are completely independent of each other, the entries in the dataset (distribution of mass) can be reproduced from the row and column totals alone, or *row and column profiles* in the terminology of CA.

5. RESULTS

5.1 *Solutions by a Classical Regression Problem Approach with the Explanatory Variables*

According to the linear regression of all the 9 *explanatory variables* in the dataset, only the intercept parameter β_0 is statistically significant. Hence, none of these variables help to predict the *response variable*, **mpg** *(miles per gallon)* (see Table 2). The residual plots in Figure 1 suggest that the residuals appear to be reasonably *normal* with constant variance, given the sample size.

TABLE 2. - *Summary Statistics for Regression Model for all Explanatory Variables*

Coefficient	Estimate	Std. Error	t Value	$\Pr(>\|t\|)$
Intercept	37.37	17.56	2.13	0.05
displacement	−0.09	0.06	−1.66	0.11
hp	−0.06	0.08	−0.74	0.47
torque	0.08	0.09	0.94	0.36
carburetor	1.20	1.31	0.91	0.37
transmission.speed	0.43	2.13	0.20	0.84
length	0.06	0.09	0.67	0.51
width	−0.24	0.27	−0.91	0.38
weight	0.00	0.00	0.09	0.93
transmission.type	−0.94	3.19	−0.30	0.77

Root MSE:	3.35
R^2:	0.74
Adj. R^2:	0.71
R:	0.86
F Statistic:	8.73 on 9 and 20 degrees of freedom
p-value:	< 0.00

However, the *variance inflation factors* (VIFs) of this model in Table 3 indicate *high dependency* between some of the *explanatory variables*; thus, it allows some

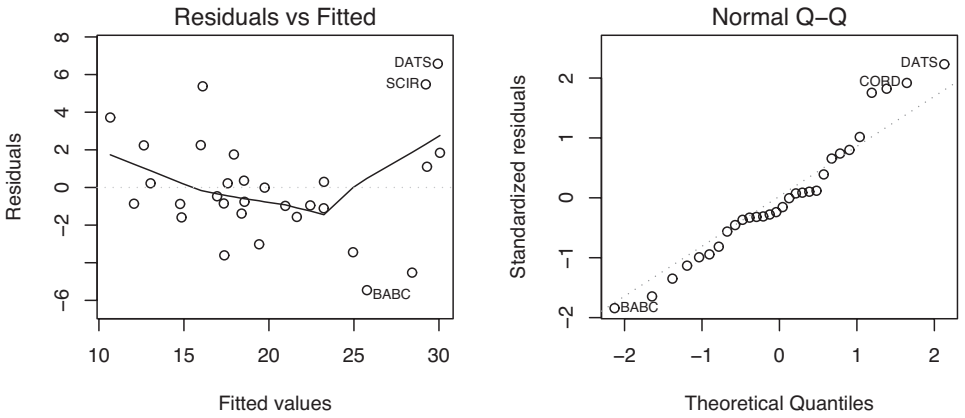

FIGURE 1. - *Residual Plots for Regression Model for all Explanatory Variables*

variables to be removed from the analysis. The variables, *displacement, torque and carburetor*, can be removed because they are dependent on *hp* of the automobile. The variable *weight* can be removed too because it is dependent on the *length* and *width*.

TABLE 3. - *Variance Inflation Factors of Regression Model for all Explanatory Variables*

Coefficient	VIF
displacement	102.61
hp	35.07
torque	126.73
carburetor	4.98
transmission.speed	4.96
length	9.83
width	5.77
weight	19.66
transmission.type	5.16

According to the new linear regression analysis in Table 4, at least one of these variables helps to predict the *response variable*, **mpg** or *miles per gallon*. The *residuals plots* in Figure 2 suggest that the residuals appear to be reasonably *normal* with constant variance, given the sample size. Although there appear to be some outliers in these plots, the standardized residuals in Table 5 suggest that they should still be included in the model.

TABLE 4. - *Summary Statistics for Regression Model after excluding Highly Dependent Explanatory Variables*

Coefficient	Estimate	Std. Error	t Value	$\Pr(>\|t\|)$
Intercept	45.02	14.41	3.13	< 0.00
hp	−0.06	0.03	−2.40	0.02
transmission.speed	1.70	1.94	0.87	0.39
length	0.03	0.08	0.44	0.67
width	−0.41	0.25	−1.63	0.12
transmission.type	1.38	3.03	0.46	0.65

Root MSE:	3.46
R^2:	0.75
Adj. R^2:	0.69
R:	0.87
F Statistic:	14.2 on 5 and 24 degrees of freedom
p-value:	< 0.00

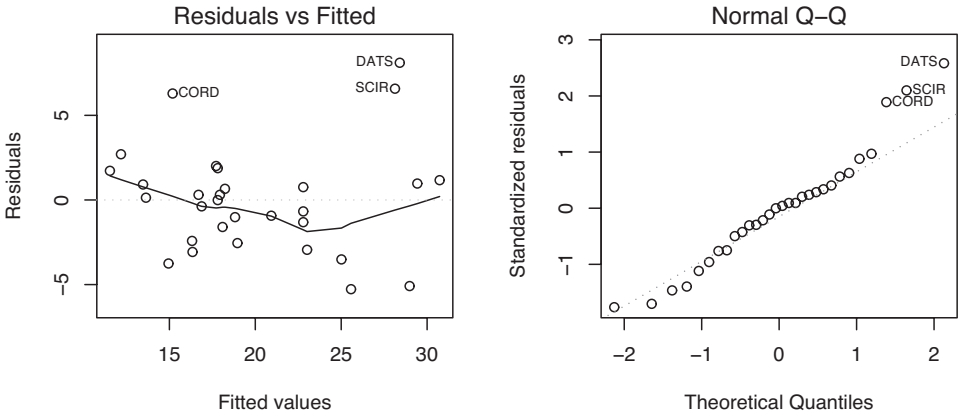

FIGURE 2. - *Residual Plots for Regression Model after excluding Highly Dependent Explanatory Variables*

TABLE 5. - *Summary Statistics for Regression Model after excluding Highly Dependent Explanatory Variables*

Obs.	Std. Resid.	Obs.	Std. Resid.	Obs.	Std. Resid.
apol	0.19	cam6a	−0.11	star	0.22
omeg	0.09	dats	2.34	cord	1.82
nova	−0.27	capr	−0.02	coe5	0.34
mona	0.09	pace	0.54	mark	0.50
dust	−0.85	babc	−1.52	celi	−1.47
jens	−1.08	gran	−0.29	char	0.58
skyh	−0.19	eldo	0.27	coug	−0.70
monz	−0.38	impe	0.78	elit	−0.89
scir	1.90	novl	−0.00	mata	0.04
cosr	0.28	vali	−0.73	corg	−0.46

The VIFs of this model in Table 6 are small (≤ 10), indicating low dependency amongst these variables.

TABLE 6. - *Variance Inflation Factors of Regression Model for all Explanatory Variables*

Coefficient	VIF
hp	3.28
transmission.speed	3.98
length	6.42
width	4.82
transmission.type	4.49

According to *Mallow's Cp* in Table 7, a model having 3 or 4 variables appears to be best regression model for the remaining variables. The best model having 3 variables contains the variables, *hp, transmission.speed* and *width*, and the best model having 4 variables, *hp, transmission.speed, width* and *transmission.type*.

TABLE 7. - *Mallow's Cp Selection Method*

	Variables				
	1	2	3	4	5
Cp Values	8.67	3.24	2.36	4.19	6.00

Selection Algorithm: Exhaustive					
Variables	hp	transmission.-speed	length	width	transmission.-type
1	*				
2	*			*	
3	*	*		*	
4	*	*		*	*
5	*	*	*	*	*

According to the PRESS statistic, the model having 3 variables is the best model. The first value (404.72) is the PRESS statistic for the model using 3 variables while the second value (462.89) is the PRESS statistic for a model using 4 variables. The model with the smaller PRESS statistic is preferred.

In Table 8, the best model to predict **mpg** *is*

$$\textbf{mpg} = 44.03 - 0.06\ hp + 2.29\ transmission.speed - 0.33\ width$$

In the last column of Table 8, the asymptotic p-value for each coefficient is given. That should be precise enough.

TABLE 8. - *Summary Statistics for Best Subset Regression Model*

Coefficient	Estimate	Std. Error	t Value	$\Pr(>\|t\|)$
Intercept	44.03	13.62	3.23	< 0.00
hp	−0.06	0.02	−2.95	0.01
transmission.speed	2.29	1.31	1.75	0.09
width	−0.33	0.17	−1.93	0.06

Root MSE:	3.35
R^2:	0.74
Adj. R^2:	0.71
R:	0.86
F Statistic:	25.12 on 3 and 26 degrees of freedom
p-value:	< 0.00

Figure 2 and Figure 3 are not the same plots although they look the same. Those two variables that were removed from the full model made little or no contribution to the predictions, and thus, little change to residual plots.

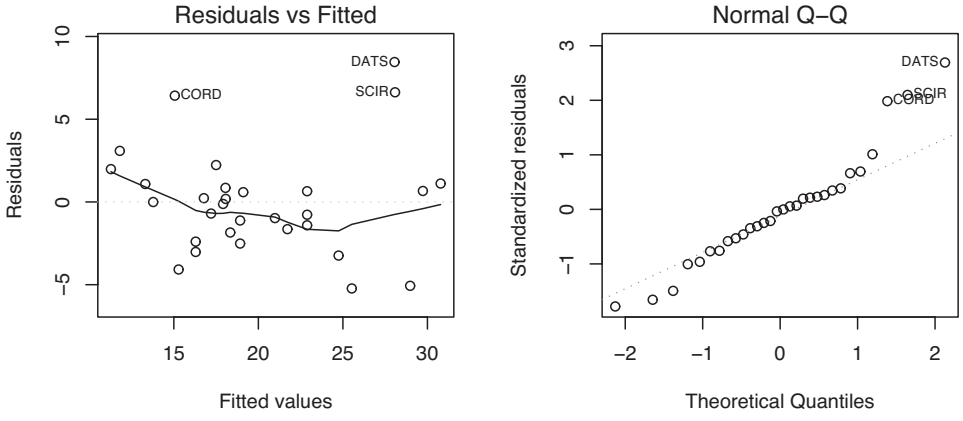

FIGURE 3. - *Residual Plots for Best Subset Regression Model*

5.2 *Preliminary Agglomerative Hierarchical Clustering of the Indicator Matrix* $\mathbf{S}_{30\times39}$

We perform a preliminary clustering of the observations with χ^2 distance (*criterion of inertia*), using an *agglomerative algorithm* which is locally robust in the sense that the lower parts of the produced dendrogram are largely independent of possible outliers. For this reason, ''classification'' is used in this Section 5.2 as a prior condensation of the data. For more details, see Lebart (1994).

Here the categorical data have been recorded as dummy variables $J = \sum_q J_q$ columns.

More precisely, we have carried out on the *complete disjunctive table* $\mathbf{S}_{30\times39}$ crossing the 30 observations (automobiles) while splitting up the 9 *explanatory variables* into 34 modalities and the *response variable* into 5 categories (*MIL*1, *MIL*2, *MIL*3, *MIL*4, *MIL*5), for a total of 39 modalities.

When analyzing Figure A.1, we see that the class 55, *the class of the* 8 *manual cars* of our basic sample, breaks away from the top of the tree. The remainder, i.e., class 58, *the class of automatic cars* is divided into 57 and 53. The histogram of the indices of level of the hierarchy indicates two ruptures, suggesting the *general tendency* which will be confirmed later by CA in Section 5.3.2, i.e., the split according to F_1: opposition between the different types of transmission speeds.

It is certainly necessary to be careful not to excessively rationalize. *Agglomerative hierarchical clustering*, founded on a sample, cannot provide an explanatory model; but it has merit in that it invites us to treat and compare the model, by proposing a way of looking at it in terms of functionality. Thus, with CA, one reduces the number of dimensions; by hierarchical clustering, one reduces the number of types.

To confirm the general tendencies of this preliminary clustering, we perform CA of the same indicator matrix $\mathbf{S}_{30\times39}$. CA of this table $\mathbf{S}_{30\times39}$ leads to five important eigenvalues (see Table 9). The first plane (represented in *Figure A*.2) accounts for 43.43% of the total variance.

TABLE 9. - *Correspondence Analysis of Table* $\mathbf{S}_{30\times39}$*: Eigenvalues*

	Eigenvalues	%	% Cumul.	Histogram
1	0.79	27.33	27.33	**********
2	0.47	16.10	43.43	******
3	0.34	11.79	55.22	****
4	0.27	9.15	64.37	***
5	0.17	6.00	70.37	**

The sequence of patterns can be observed along a succession of *response variables*, going in the parabolic direction known as the *Guttman effect* (a typical structure we usually find in ordered categorical data): from the strongest (*MIL*5) to lowest fuel consumption (*MIL*1). This sequence is typical of a *hierarchical structure*: the non-zero coordinates on each principal axis oppose mainly two groups of automobiles: cars with automatic transmission and cars with manual transmission.

The first axis, for instance, according to factor projections, contributions and correlations, is the axis of *manual transmission* (*scir*, *cosr*, *dats*, *capr*, *babc*, *coe*5, *celi*) with very pronounced negative coordinates except *dust*.

These manual cars are associated with the response category *MIL*5 and the explanatory categories *typ*2, *dis*1, *hor*1, *tor*1, *nts*2, *ovl*1, *wid*1, *wei*1. More precisely, the weaker displacement (cubic in.), horsepower (ft-lb), torque (ft-lb), overall length (in.), width (in.), weight (lb) are, the more the number of transmission speeds is raised, and the more consumption of gas will increase for the cars of manual transmission.

Axis 2 opposes an intermediate class $(MIL2, MIL3)$ to the two extreme classes $(MIL1, MIL5)$ of consumption of gasoline. However, plane projections do not allow the definition of classes with as much finesse and certainty: one could resort here to using the tools to interpret and to better label the tree resulting from hierarchical clustering. More details for the interpretive tools are furnished by the programs *Facor* or *Vacor*; see Benzécri (1992) and Murtagh (2005).

5.3 *Regression Analysis used with MCA*

This approach projects explanatory variables on a lower dimensional space that almost estimates the response variable *y*. This methodology, divided into two parts, is at the heart of MCA, studying the regression problem between a response variable and a set of categorical predictor variables:

1. Interpreting relationships between the response variable *y* and the predictor variables from different axes (planes 1-2, 1-3, etc.) of the previous CA which yields two clouds of points, namely the cloud of modalities, and the cloud of individuals (or equivalently-weighted response patterns).
2. Regression analysis where we consider explanatory variables as factor coordinates (on the observations) provided from the previous CA.

5.3.1. CA of the Burt Table $\mathbf{B}_{11\times 11}$

We believe it to be more useful to explore another approach here: to subject the 3 questions (2 explanatory and 1 response variables) of Table 1 to a complete disjunctive form (in the format of an indicator matrix); from which is built a Burt matrix (Burt table).

The 3 variables are divided into 11 categories: 5 *for miles/gallon*, 4 *for displacement and* 2 *for the type of transmission (typ1: automatic; typ2: manual)*.

The *''functional model''* which one produces initially is that of the *''continuous correspondences''*; by dividing the explanatory variables into a great number of categories, there would be for first coordinate, the function, of general form, best correlated with the quantity to explain. For more details, see Benzécri (1992,p. 392); Lebart *et al.* (1997,p. 108), Cazes (1977,1997) and de Tibeiro (1997).

CA of the *disjunctive table*, is also called an *indicator matrix*, crossing the 30 observations and $5 + 4 + 2 = 11$ modalities. That is, MCA, which yields two clouds of points, namely the cloud of *11* categories, and the cloud of *30* observations. In numerical terms, each cloud is defined by a table of principal coordinates, where, for each axis, the weighted average of the squares of the principal coordinates is equal to the eigenvalue associated with the axis. For more details, see Murtagh (2005), Le Roux and Rouanet (2004) and Greenacre (1991).

CA of the Burt table is now discussed. The percentages of inertia explained by the top-ranked five factors are 55.28%, 19.12%, 11.70%, 7.49% and 4.72%. Axis 1 therefore accounts for more than one half of the total inertia of the cloud. We have displayed the principal (1,2) plane in Figure A.3, representing 74.40% of the total inertia, in which the initial character of the first factor appears.

The first factor (F_1), clearly stands apart from the one that follows it in the table of the eigenvalues. It is *a factor of general level* (the first eigenvalue which is associated with the first axis is almost double the two subsequent ones). This factor indicates the degree of separation of the different automobiles according to the gasoline mileage performance system.

We are thus lead to deal with this first axis as a *new artificial variable*, providing *a mode of ''classification''* of the observations taken from our basic sample. The significance of the known factor will be illustrated by exogeneous information provided from explanatory variables. The classic linear regression would not be precise enough here if one applied it to the primary data. It was necessary to create, a ''new explanatory variable'' which incorporates into the study the explanatory variable as it creates gas consumption (mileage).

F_1 is a *general attitude* which necessarily detaches itself from the attitudes based on particular facts. We note also that the first three factors explain a little bit more than 86.10% of the dispersion of observed values. We will admit that the essential part of structural links between the data is contained in the space of the three first dimensions.

This first axis is produced when contrasting two types of vehicle transmission (*manual transmission, typ2: automatic transmission, typ1*). On the side $F_1 < 0$, the cars of *manual transmission* are associated to a *very strong fuel consumption* $(MIL5)$ and a *weak rate of displacement* $(dis1)$. On the other hand $(F_1 > 0)$, the cars with *automatic transmission* are related to a *low fuel consumption* $(MIL2, MIL1, MIL3)$ and a strong enough rate of displacement $(dis3, dis4)$.

Axis 2 confirms the split between gas consumption $(MIL3, MIL1)$ and displacement $(dis2, dis4)$. CA therefore has created a synthetic variable, a *factor of general level*, an expected indicator: the contrast between the superior categories and inferior categories at the response variable level $(miles)$ and the explained variable level (displacement (cubic in.) and type of transmission).

Thus, by diving up the explanatory variables into a great number of categories, one would have for *first factor*, the best correlated general form function with the response variable y, linear combination of the factors $F_\alpha(i)$ resulting from CA. For more details, see Benzécri (1992), Cazes (1977, 1978), Lebart *et al.* (1997) and de Tibeiro (1997).

The 1×2 plane in Figure A.3 reveals an excellent ''discrimination'', a clear separation between the cars of automatic transmission and the cars of manual transmission. What confirms the opposition evoked above in Section 5.2 between class 55 [of the cars with manual transmission, $(F_1 > 0)$] and class 58 [of the cars with automatic transmission, $(F_1 < 0)$]. This similarity of the results appears acceptable to us even if CA of the Burt table $\mathbf{B}_{11\times 11}$ is related only to 3 variables [Miles/gallon, Displacement (cubic in.) and Type of transmission].

5.3.2. CA of the Table $\mathbf{C}_{5\times 34}$

According to Steps 1 and 2 of Section 3, we have created the table $\mathbf{C}_{5\times 34} = \mathbf{T}^t_{5\times 30}\mathbf{S}_{30\times 34}$ where $\mathbf{T}^t_{5\times 30}$ is the transpose of $\mathbf{T}_{30\times 5}$ (the complete disjunctive form associated with the response variable) and $\mathbf{S}_{30\times 34}$: the complete disjunctive form associated with the 9 explanatory variables.

We have displayed the (1,2) plane in Figure A.4 which is almost sufficient for the interpretation. It shows the set of all the categories of the response variable y (mileage) spread out on a *parabolic crescent*. It is an index of a steep gradient within the data: these are arranged according to a series which is patently obvious, not on the axis 1, but in the plane 1×2.

We consider the 1×2 plane, in which the initial character of the first factor appears: the 5 categories of the response variable and the 34 modalities associated with the explanatory variables, reveals an *excellent discrimination* of the two types of transmission (manual and automatic). There is a succession, according to the first axis, going in the parabolic direction of *Guttman effect*: from the strongest $(MIL5)$ to lowest fuel consumption $(MIL1)$.

In practice, the set of the selected explanatory variables X should be enough to approximatively reconstruct the response variable y. We believe, it to be an accurate expression of the relative importance of the factors. The percentage of inertia explained by the top-ranked factors of this table $C_{5\times 34}$ are

$$\tau_1 = 61.40\%; \tau_2 = 23.78\%; \tau_3 = 9.29\%; \tau_4 = 5.54\%$$

Therefore, one notices the preponderance of the first factor (axis 1), with 61.40% of inertia. This factor clearly stands apart from the one that follows it in the table of the eigenvalues $(\tau_1 = 61.40\%; \tau_2 = 23.78\%)$, accounts for more than half of the inertia of the cloud. It is also *a factor of general level*: a general attitude which necessarily detaches itself from the attitudes based on particular facts.

On *axis* 1 (according to factor projections, contributions and correlations), we see that the first dimension contrasts with the *strong mileage* $(MIL5)$ (positive part) and the *low mileage* (*MIL*1 and *MIL*2) (negative part). The response category $(MIL5)$ is associated with the explanatory categories *dis*1, *hor*1, *tor*1, *nts*2, *ovl*1, *wid*1, *wei*1 and *typ*2 whose correlations are very good. The response categories (*MIL*1 and *MIL*2) are associated with the other explanatory categories, particularly *dis*3, *car*2, *wid*4, *hor*3, *ovl*5, *wei*3, *wei*5, *nts*1 and *typ*1.

On *axis* 2, $(MIL1)$ is the opposite of all the *response variables* except $(MIL5)$ whose contribution and correlation are negligible. More precisely, on this axis, $(MIL1)$ is associated with the following explanatory variables: *dis*4, *hor*5, *tor*5, *ovl*5, *wid*4 and *wei*5.

5.3.3. ''Visualized Regression'' and CA of the ''Regression Table'' $C_{35\times34}$

According to Step 3 in Section 3, we perform an estimation of the response variable from the *''Regression Table''*. One adds up the *indicator matrix* $S_{30\times34}$ (supplementary rows or vectors of description in (0,1) of all the individuals of the basic sample) as supplementary to $C_{5\times34}$, while projecting on the first four (non trivial) factorial axes found the profiles of the rows e of table S. Each modality of the response variable is regarded here as a numeric variable, that one seeks to express in linear combination of the data variables, replaced here by the factors resulting from CA of the table in (0,1) or of *the table of ''regression''*.

In this Figure A.5, we consider the same results evoked above in Figure A.4 where we project moreover additional categories relating to the marks of the cars. For the first group of variables on the positive side $(F_1 \geq 0)$ of Figure A.5, the strong categories $(MIL5, MIL4)$ are associated with the type of *manual transmission* (*typ2*).

Of course, the manual cars projected in this group are *capr, babc, scir, monz, skyh, dust, celi*. See Agglomerative Hierarchical Clustering on Figure A.1. This ''visualized regression'' confirms to some extent the result already evoked in preliminary clustering of the indicator matrix **S** (Section 5.2). See Agglomerative Hierarchical Clustering in Figure A.1.

For a second group of variables $(F_1 \leq 0)$, it is the contrary. The weak ones and average modalities of mileage $(MIL1, MIL2, MIL3)$ are associated with the *automatic type of transmission* (*typ1*). Practically all cars with automatic transmission of the basic sample are projected there: *eldo, jens, coug, cord, char, nova, apol, gran, omeg, cama, mona, vali, corv, nova,...* See Agglomerative Hierarchical Clustering in Figure A.1.

Between the two groups, *the separation is almost perfect as* evoked previously in Section 5.3.1. We find the same *Guttman Effect* in the plane (1,3) of Figure A.6 which emphasizes clearly the curve of the modalities with the form of an *S*. Agglomerative Hierarchical Clustering of the two sets in correspondence confirms fully the interpretation of factors 1 and 2. We connected the categories of the same parameter (or variable) by a polygonal line, thus highlighting a gradient of the function of the type of transmission in the direction of axis 1.

The cloud of *supplementary ''individuals''* is projected in the plane (1,2) with for zone of maximum density, a crescent framing the curve of the modalities of the response variable.

6. CONCLUDING REMARKS

Preliminary results indicate that there is a difference between the regression lines relating displacement to mileage for automatic (*typ1*) and manual driven vehicles (*typ2*). We realize that regression analysis applied to the original Table 1 is equivalent to MCA of the data table $C_{35\times34}$ which can be regarded as the estimation of the response variables categories by means of the explanatory variables.

It is realized here that in practice the set of these selected explanatory variables X is enough to reconstitute approximatively the response variable y.

ACKNOWLEDGEMENTS

The first author is grateful to Paul Nguyen, doctorate student at the Department of Statistical & Actuarial Sciences of the The University of Western Ontario (Canada) for the assistance in the computation and the analysis of the results.

RIASSUNTO

Lo studio della dipendenza tra variabili di solito è affrontato con la regressione multipla, ma talvolta la struttura dei dati è complessa e richiede un approccio più articolato che deve anche tener conto delle potenzialità grafica delle tecniche multivariate. L'articolo integra l'analisi della regressione multipla con l'Analisi delle Corrispondenze Multiple. Viene altresì considerata anche l'analisi di classificazione automatica che si accompagna spesso all'Analisi delle Corrispondenze Multiple. Infine, sono evidenziati con un data set noto i vantaggi di questo approccio integrato.

APPENDIX

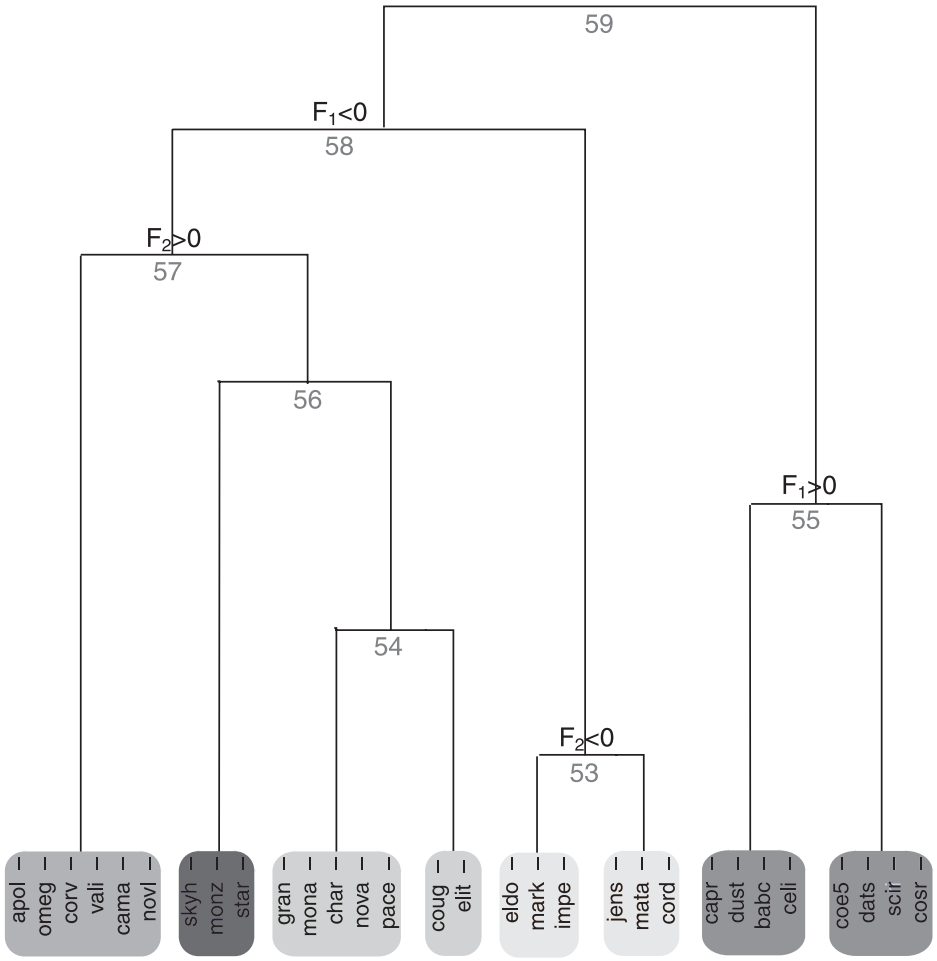

FIGURE A.1 - *Dendrogram Associated with the Table* $\mathbf{S}_{30\times39}$

Preliminary **Hierarchical Agglomerative Clustering** carried out on the *indicator matrix* $\mathbf{S}_{30\times39}$, crossing the 30 *automobiles*, while expanding the 9 explanatory variables into 34 modalities and the response variable into 5 categories, for a total of 39 *modalities*.

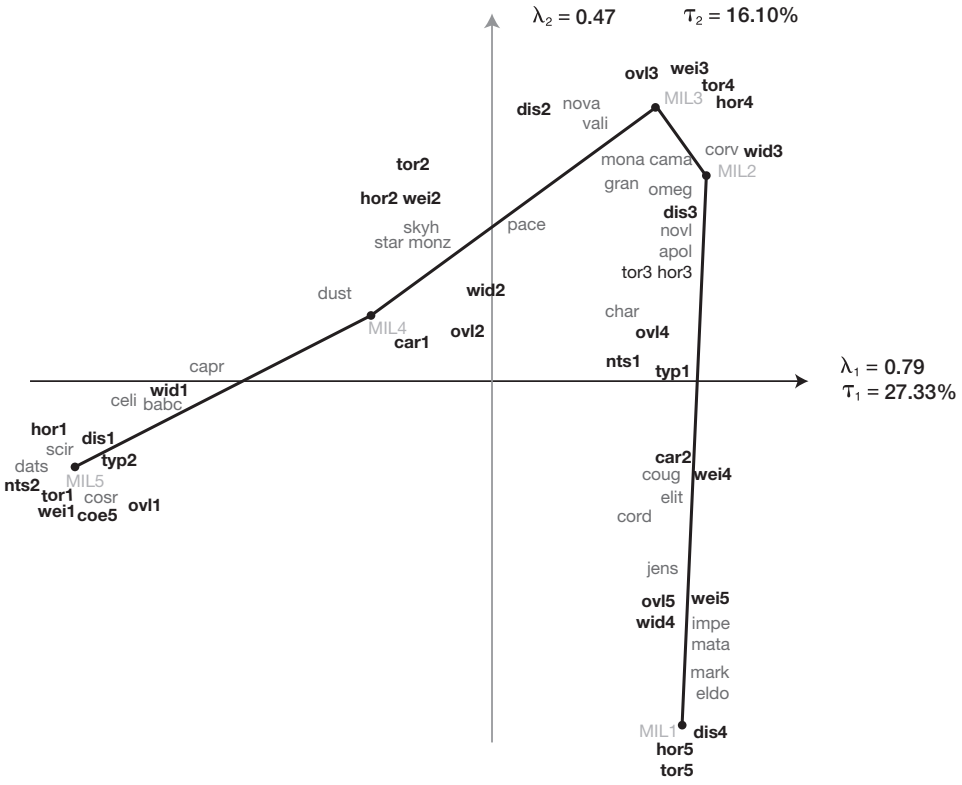

FIGURE A.2 - *Correspondence Analysis Factor Map*

Plane spanned by axes 1 and 2: **CA of** *the table* $\mathbf{S}_{30\times39}$ crossing the 30 observations (automobiles) while splitting up the 9 *explanatory variables* into 34 modalities and the *response variable* into 5 categories (*MIL*1, *MIL*2, *MIL*3, *MIL*4, *MIL*5), for a total of 39 modalities.

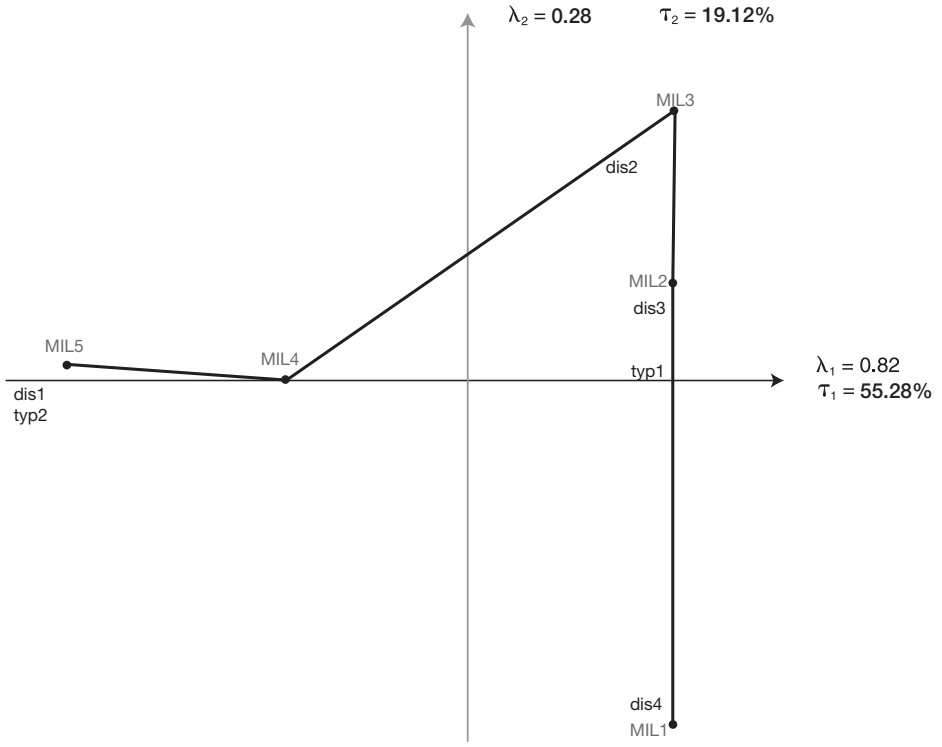

FIGURE A.3 - *Correspondence Analysis Factor Map*

Plane spanned by axes 1 and 2: **CA of** *the Burt's table* $\mathbf{B}_{11\times11}$ generated by *3 variables (Miles/gallon, Displacement (cubic in.) and Type of transmission)* of *Table* 1.

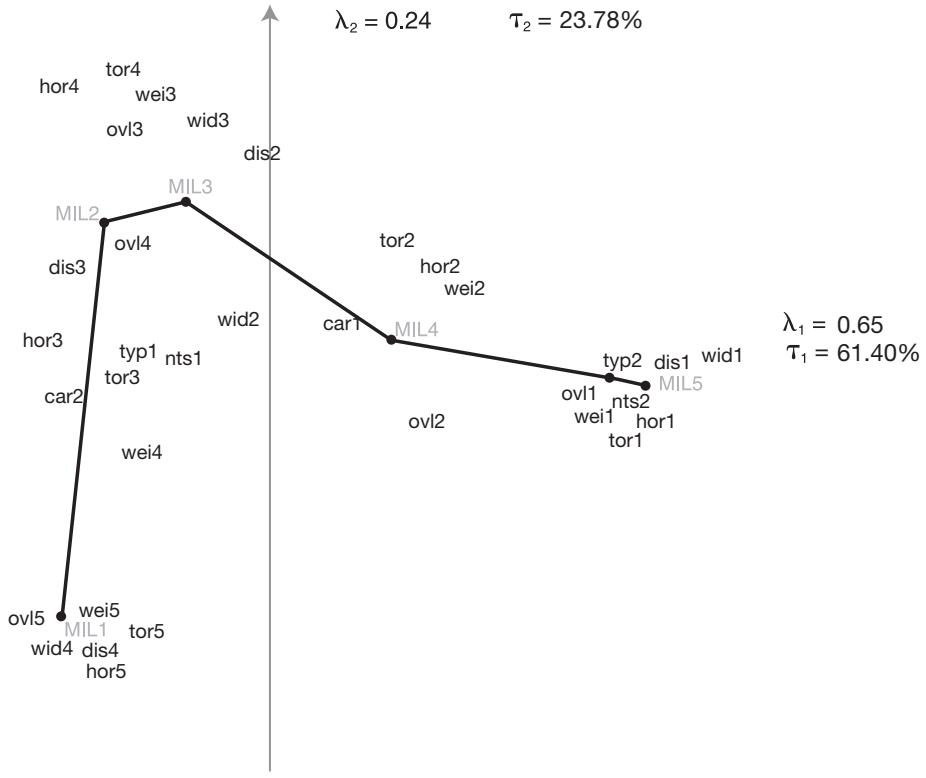

FIGURE A.4 - *Correspondence Analysis Factor Map*

Plane spanned by axes 1 and 2: **CA of** *the table* $\mathbf{C}_{5\times34} = \mathbf{T}^{t}_{5\times30}\mathbf{S}_{30\times34}$ according to Steps 1 and 2 of Section 3. $\mathbf{T}_{30\times5}$: the *complete disjunctive form* associated with the *response variable* and $\mathbf{S}_{30\times34}$: the *complete disjunctive form* associated with the 9 *explanatory variables*.

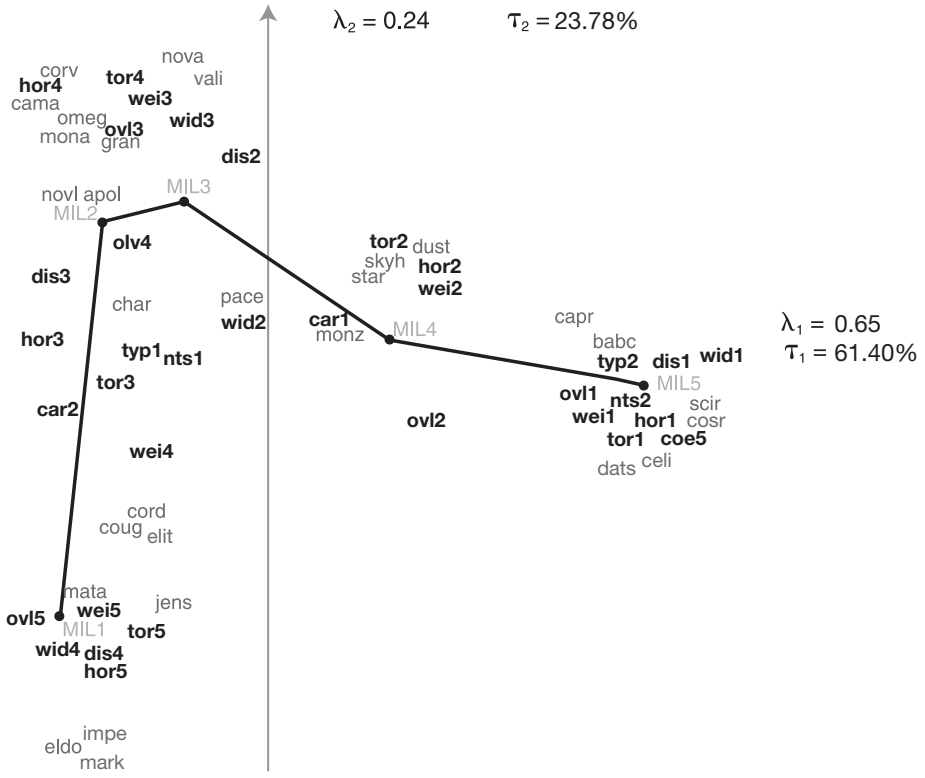

FIGURE A.5 - *Correspondence Analysis Factor Map*

Plane spanned by axes 1 and 2: **CA of** *the ''Regression Table''* $\mathbf{C}_{35\times34}$ according to Step 3 in Section 3.

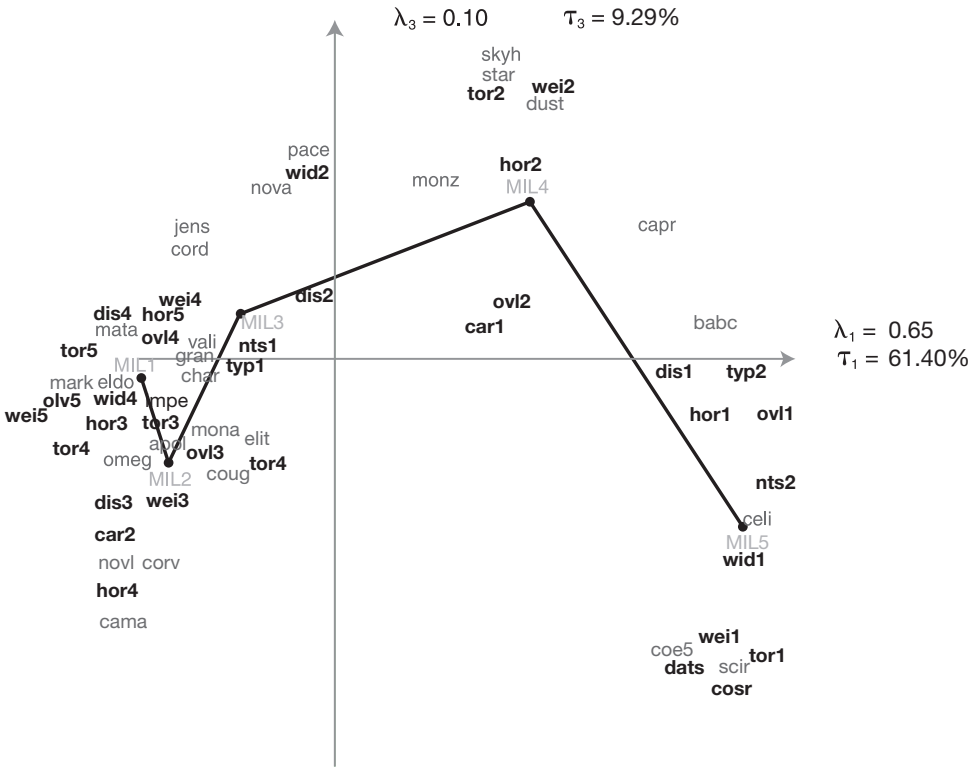

FIGURE A.6 - *Correspondence Analysis Factor Map*

Plane spanned by axes 1 and 3: **CA of** *the ''Regression Table''* $\mathbf{C}_{35\times34}$ according to Step 3 in Section 3.

REFERENCES

Benzécri J.-P. et Coll. (1973). *L'Analyse des Données. Tome I: Taxinomie. Tome II: Analyse des Correspondances*. Dunod, Paris.

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

Benzécri, J.-P. (1997a). Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Données*, **2**, 191-198.

Benzécri, J.-P. (1997b). Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Données*, **2**, 209-218.

Brenot J. (1977). *Contributions à la pratique du modèle linéaire: Qualité, protection et estimation biaisée*. Thèse de 3ème cycle, Paris VI.

Cazes P. (1975). Protection de la régression par utilisation des contraintes linéaires et non linéaires. *Revue de Statistique Appliquée*, **23**(3) 37-57.

Cazes P. (1977). *L'école d'été du CNRS sur l'Analyse des Données, laboratoire du Pr. J.-P. Benzécri*. Université Pierre et Marie Curie, Paris VI.

Cazes P. (1978). Méthodes de régression III. L'Analyse des donnés. *Les Cahiers de l'Analyse des Données*, **3**, 385-391.

Cazes P. (1990). Codage d'une variable continue en vue de l'analyse des correspondances. *Revue de Statistique Appliquée*, **38**(3) 33-51.

Cazes P. (1997). Adaptation de la régression PLS au cas de la régression après l'analyse des correspondances multiples. *Revue de Statistique Appliquée*, **45**(21) 89-99.

D'Ambra L., Amenta P., Gallo M. (2005). Dimensionality reduction methods. *Metodoloski zvezki*, **2**(1) 115-123.

de Tibeiro J. (1997). Consommation d'électricité sous un climat extrême: Estimation en fonction de la date et de la température. *Les Cahiers de l'Analyse des Données*, **22**(2) 199-210.

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

Greenacre M.J. (1991). Interpreting multiple correspondence analysis. *Applied Stochastic Models and Data Analysis*, **7**, 195-210.

Jambu M., Lebeaux M.O. (1983). *Cluster Analysis and Data Analysis*. North-Holland Publishing Co, Amsterdam.

Juan J. (1982). Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques. *Les Cahiers de l'Analyse des Données*, **7**, 219-225.

Lebart L. (1994). Complementary use of correspondence analysis and cluster analysis. In M.J. Greenacre and J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and and Applications* (pp. 162-178). London: Academic Press.

Lebart L., Morineau A., Piron M. (1997). *Statistique exploratoire multidimensionnelle* (2ème édition). Dunod, Paris.

Lebart L., Morineau A., Warwick K. (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley, New York.

Le Roux B., Rouanet H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Kluwer, Dordrecht.

McQuitty L.L. (1966). Similarity analysis by reciprocal pairs of discrete and continuous data. *Educational and Psychological Measurement*, **26**, 825-831.

Montgomery D.C., Peck E.A., Vining G.G. (2001). *Introduction to Linear Regression Analysis* (3rd edition). John Wiley & Sons Inc, New York.

Murtagh F. (2002). Clustering in massive data sets. In J. Abello, P.M. Pardalos and M.G.C. Resende (Eds.), *Handbook of Massive Data Sets* (pp. 501-543). Norwell, Kluwer Academic Publishers.

Murtagh F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall/CRC, Boca Raton.

Tenenhaus M. (1998). *La Régression PLS, Théorie et Pratique*. Technip, Paris.