COMMUNITY ECOLOGY 8(1): 25-30, 2007
*1585-8553/$20.00 © Akadémiai Kiadó, Budapest*
*DOI: 10.1556/ComEc.8.2007.1.4*

# Comparison of single and complete linkage clustering with the hierarchical factor classification of variables

S. Camiz[1] and V. D. Pillar[2]

[1] *Corresponding author. Dipartimento di Matematica Guido Castelnuovo, Università di Roma La Sapienza, Piazzale Aldo Moro, 2, I-00185 Roma, Italia. E-mail: sergio@camiz.net*
[2] *Departamento de Ecologia, Universidade Federal de Rio Grande do Sul, 91540-000, Porto Alegre, RS Brazil. E-mail: vpillar@ufrgs.br*

**Keywords**: Classification of variables, Comparison of methods, Hierarchical classification, Principal Components Analysis, Randomization tests, Simulated correlation matrices.

**Abstract**: We assess the performance of a new clustering method for Hierarchical Factor Classification of variables, which is based on the evaluation of the least differences among representative variables of groups, as defined by a set of two-dimensional Principal Components Analysis. As an additional feature the method gives at each step a principal plane where both grouped variables and units, as seen only by these variables, can be projected. We compare the method results with both single and complete linkage clustering, applied to simulated data with known correlation structure and we evaluate the results with a coherence measure based on the entropy between the expected partitions and those found by the methods. We found that the Hierarchical Factor Classification method performed as good as, and in some cases better than, both single and complete linkage clustering in detecting the known group structures in simulated data, with the advantage that the groups of variables and the units can be viewed on principal planes where usual interpretations apply.

**Abbreviations:** *HFC* – Hierarchical Factor Classification, *PCA* – Principal Components Analysis

## Introduction

In a previous paper, Camiz et al. (2006) introduced the Hierarchical Factor Classification (*HFC*) of variables which builds a hierarchy on a set of variables with a recursive procedure. In each step, the algorithm computes a centered non-standardized *PCA* on pairs of variables, each representative of a group. Then it chooses for merging the two groups for which the second *PCA* eigenvalue is minimum and considers the first principal component as representative of the newly formed group. By centered non-standardized we mean that *PCA* uses the covariance matrix rather than the correlation matrix of the variables.

This method is one of the very few specifically designed for clustering variables. Several advantages are attributed to it in comparison to the others, especially its ability to provide principal planes associated with each node of the hierarchy, where both variables and units can be represented, as in a common *PCA*. On these planes, all variables gathered in the group are represented as well as all units as seen only by these variables. Indeed, in this case the positions of points on the plane depend only on the two involved representative variables, thus only on the variables of the group. In fact, the representative variables are linear combinations of the variables belonging to the represented group. This may be used to understand the influence of the variables of each group on the point pattern on the principal planes. As a result, the interpretation of the principal components is straightforward, since the first one represents what the gathered variables have in common, whereas the second one shows their differences. In addition, the gathering of variables regardless of the sign of their mutual correlation gives groups that have the form of dipoles, say two subgroups opposed in their meaning. This makes easier the interpretation of the results for the end user.

In this paper, we compare the performance of this method with other hierarchical classification mehods. In particular, we shall deal with both single and complete linkage methods (Anderberg, 1973, Legendre and Legendre, 1998, Podani, 2000) based on the correlation matrix.

The comparison will be based on simulated data with known group structure.

### Hierarchical classification of variables

In the literature, the classification of variables received little attention and very few methods are currently available, especially hierarchical methods are lacking. Therefore, the same methods used for classifying units are applied to variables in the software programs currently available. Indeed, only in *SAS* (*SAS* Institute, 1999) the procedure *VARCLUS* is proposed, which is a divisive algorithm that at each step tries to define groups of variables as unidimensional as possible. Other methods that appear interesting, such as Lermann (1991) and Vigneau et al. (2006) are based on different principles and are not easily available for a standard use.

For this reason, we compare here *HFC* with the two simplest hierarchical methods that can be easily found or implemented: the single and complete linkage. These methods can be used regardless of the kind of objects that are being classified and need only that a resemblance (association) matrix exists measuring the degree of similarity or dissimilarity among the objects. Indeed, on this basis two objects will be considered most similar if their similarity is maximum or their dissimilarity is minimum.

In order to build a hierarchy of variables, the abovementioned algorithms are agglomerative: at the beginning each variable forms a group and the association matrix contains the defined pairwise relations among variables; then:

1) each agglomerative step joins the two existing groups that optimise the chosen objective function;

2) the association matrix is redefined according to the new group structure resulting from step 1;

3) the process is repeated until all variables are joined in one group.

If $n$ is the number of variables to classify, there will be $n$-1 clustering steps. The clustering process will indicate $n$-1 possible partitions and the decision on which partition to take is left to the user (see Milligan and Cooper 1985, for a cricital evaluation).

The algorithms differ in both the objective function to be optimized and the criterion for redefinition of intergroup association (step 2). It is questionable how to apply methods other than single and complete linkage to variables, since it does not make sense to calculate average covariances or correlations, or Euclidean distance among variables, as required e.g., by Ward's (1963) method. In addition, since they are a kind of "extreme" methods, the comparison with other monotone clustering techniques seemed unnecessary. So, we limited our attention to single and complete linkage which do not recalculate the associations among the formed groups but search at each step either the minimum or the maximum. In single linkage clustering (Florek et al., 1951, Sneath, 1957, Gordon, 1999), the association between two groups will be the one between the most similar variables in these groups. In complete linkage clustering (Sørensen, 1948, Lance and Williams, 1967, Gordon, 1999), it will be the association between the most dissimilar variables in these groups: this would guarantee, at each step, that each cluster would not contain variables whose correlation is less than the current fusion level. In both cases, it is wiser not to consider the sign of the correlation, since two variables with high negative correlation may reflect similar ecological processes, though in opposite directions, thus they should be associated at a higher level than two independent variables with near-zero correlation.

### Hierarchical factor classification

We refer to Camiz et al. (2006) for a detailed description of the *HFC* method, as applied to quantitative (ratio scale) variables. Here, we briefly recall its basis, which may be summarized as follows:

1) At the outset, variables are standardized and each variable is considered *representative* of the singleton group composed by itself. Then, the recursive algorithm is based on the following steps:

2) All pairs of existing groups are compared through their representative variables: each pair of representative variables is submitted to a *centered non-standardized PCA*, i.e., the *PCA* of their 2 × 2 covariance matrix;

3) The pair of representative variables whose second *PCA* eigenvalue is minimum is selected;

4) The two groups of variables corresponding to the selected pair of representative variables are merged in a group that becomes a new node of the hierarchy;

5) The first principal component of this *PCA*, i.e., the set of coordinates of units on the first principal axis, is chosen as representative variable of the new node;

6) The second eigenvalue of this *PCA* is chosen as fusion level of the new node in the hierarchy.

The steps (2)...(6) are repeated $n$-1 times, leading to a complete hierarchical classification of the variables.

We remind here that in the case of standardized variables covariance equals correlation, so that if the compari-

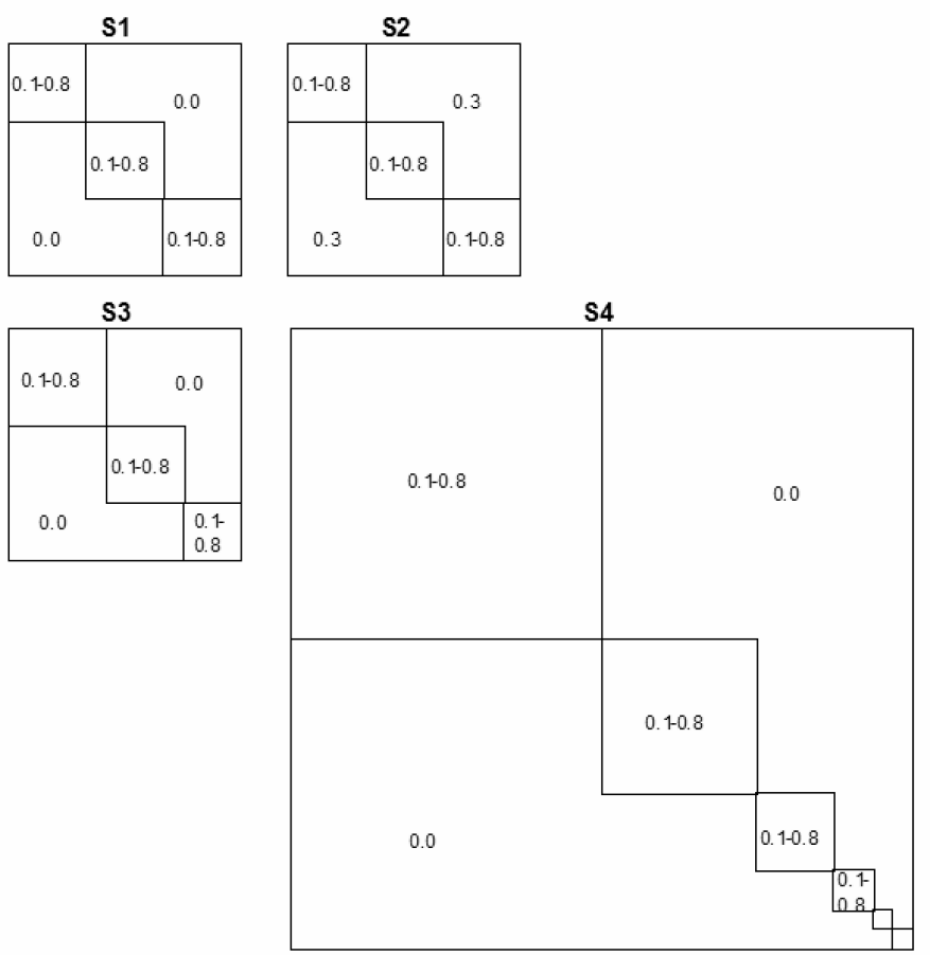


**Figure 1.** The four structures of the correlation matrices used in the experiments with different simulated data.

son is done between two original variables an ordinary *PCA* results, whereas for all other comparisons the results will be different: in particular, the trace of the matrix will be larger than 2. As a consequence, the range of the first eigenvalue is unpredictable and depends on the number of variables in the group and the agglomeration process, whereas the second one is non-decreasing along the process. According to the *PCA* logic, both eigenvalues represent a variance; the first eigenvalue is the variation common to the two merging groups, whereas the second eigenvalue is the amount of variance that is not shared. Thus, the method is akin to Ward's (1963) clustering method and it is natural to choose the second eigenvalue as the fusion level of the hierarchy.

With this method, each group may have the form of a *dipole*, since the sign of the covariance between the concerned variables has no influence on the sign of the principal components. Therefore, the variables of a node may form a dipole of two groups opposed to each other in the direction of its representative variable, according to the sign of their pairwise correlation. This is not a drawback of the method, but rather a correct idea of aggregation, since the sign of the correlation depends on how each variable is measured and not on its relation with the others.

## Experiments

To evaluate the behaviour of *HFC* in comparison with single and complete linkage clustering, we used simulated data sets with known group structure. This structure was defined by the number of variables and by the different levels of correlation between variables in the same group and in different groups.

We started by defining four types of matrices, *S1, S2, S3,* and *S4,* composed of correlations between variables gathered in groups (Fig. 1). For each type, eight matrices were defined, with the correlation between variables of the same group ranging from 0.1 to 0.8 by steps of 0.1. The types *S1*, *S2*, and *S3* all refer to 12 variables struc-

tured in three groups with different correlation levels between groups. In the matrices of type *S1*, each group contains 4 variables and the correlation between variables of different groups is 0. In the matrices of type *S2*, the structure is the same, but the correlation level between groups is 0.3. In matrices of type *S3*, the three groups contain 5, 4 and 3 variables and the correlation between groups is 0. The matrices of type S4 contain 32 variables, with the correlation structured in 6 groups containing 16, 8, 4, 2, 1 and 1 variables, with the correlation between groups equal to 0.

On the basis of these 32 matrices, we built 32 simulated data tables with a given correlation matrix. The generating procedure is that described by Ganeshanandam and Krzanowski (1990), modified by Peres-Neto and Jackson (2001), and used by Pillar (1999). Basically, it consists in a Cholesky decomposition of the specified correlation matrix **C** in the product of a triangular matrix by its transpose: **C = L'L.** If **L** is left multiplied by a unitary matrix **U** (that is such that **U'U = I**), the searched simulated data table **S = UL** is obtained, since its correlation matrix is **S'S = L'U'UL = L'L = C**. In our case, **C** was fixed as said and as **U** we used a unit matrix of standardized coordinates obtained by an ordinary *PCA* of randomly generated data tables of 1000 units and either 12 or 32 variables.

Indeed, for all 32 simulated data sets with 1000 units, the correlation matrix was exactly the one originally defined. Then, 10 samples with 30 sampling units each were taken at random from each data set. The three clustering methods were applied on all 320 samples (10 samples × 32 simulated data sets).

In order to test how well the methods could recover the expected groups in each data set, the partitions obtained by the various methods - 3 groups in *S1, S2*, and *S3*, and 6 groups in *S4* - were compared to those expected according to the pre-defined correlation structure of the simulated data. Note that we did not consider the problem of finding the optimal partition level that could lead to the identification of the expected structure, which is still an open question. The agreement was measured by the information-based *coherence coefficient* (Orlóci, 1991) computed on the contingency tables crossing the partitions obtained with the expected one. The coherence coefficient is given by:

$$\rho_{ik} = \sqrt{1 - \left( \frac{H_{i+k} - H_{ik}}{H_{i+k}} \right)^2}$$

where $H_{ik}$ is the mutual entropy and $H_{i+k}$ is the joint entropy of partitions $i$ and $k$, both of order one. In this,

$$H_{ik} = \sum_{j=1}^{s_i} \sum_{h=1}^{s_k} p_{jh} \log \left( \frac{p_{jh}}{p_{j.} p_{.h}} \right)$$

where $p_{jh}$ is the joint frequency for groups $j$ and $h$ in partitions $i$ and $k$ with $s_i$ and $s_k$ groups respectively, and $p_{j.}$ and $p_{.h}$ are the frequencies of groups $j$ and $h$ respectively, $H_{i+k} = H_{ii} + H_{kk} - H_{ik}$, where $H_{ii}$ and $H_{kk}$ are the Shannon entropies of each partition.

In order to test the statistical significance of differences among the clustering methods, an *ANOVA* was performed on the sums of squares of the coherence coefficients. The tests were based on their empirical probability distribution, obtained by a randomization test after 1000 random permutations of the coherence coefficients (Pillar and Orlóci 1996).

For data simulation and *HFC* we used a program specifically developed, whereas for the other cluster analyses and randomisation testing the *MULTIV* package was used. *HFC* is now implemented in *MULTIV* (Pillar 2006).

## Simulation results

In Figure 2, the mean and the standard deviation of the coherence coefficient between the expected partitions and those observed are reported. Mean and standard deviation are calculated on the ten replicates of any combination of correlation within groups, clustering technique, and type of matrix group structure. In each diagram, the coherence coefficient is represented according to the increase of the correlation within groups (strength of group structure), and the graphics are arranged according to the type of matrix (in row) and the clustering method (in column) considered. In all cases corresponding to these graphics, as the correlation level within the simulated groups of variables increased, all clustering methods of variables were able to reveal partitions in perfect agreement with those expected. Therefore, when groups are fuzzier (low within-group correlation levels) all methods behave worse than when they are sharper. In data sets with a sharper type of correlation structure (*S1* and *S3*), the perfect agreement was reached at a lower correlation level than the others (*S2* and *S4*). The three clustering methods did not differ in their performance (agreement with expected partition), except in the case of not so clearly defined groups in *S2* having within-group correlation of 0.4 ($P = 0.048$), and in a lesser extent in *S1* and *S3* again with within-group correlation of 0.4 ($P = 0.067$, $P = 0.058$ respectively); no difference results for *S4*; see Table 1. In these cases, pairwise contrasts indicated that the performance of *HFC* was better than both complete and single linkage in *S2* and better only than single linkage in *S1* and

**Table 1.** The coherence coefficient of the cross classification table between groups of variables indicated by three clustering methods (*HFC*, complete linkage and single linkage) and expected groups in simulated data sets generated with known correlation structure between variables. Here only the values of the correlation matrices belonging to the four types of matrices with within group correlations of 0.4 are reported, together with the probability associated by the analysis of variance based on 10 data sets generated for each combination of correlation structure.

| | S1_R4 | S2_R4 | S3_R4 | S4_R4 |
|---|---|---|---|---|
| HFC | 0.97 | 0.72 | 0.97 | 0.89 |
| Complete Linkage | 0.95 | 0.60 | 0.90 | 0.89 |
| Single Linkage | 0.92 | 0.59 | 0.87 | 0.87 |
| Probability | 0.067 | 0.048 | 0.058 | 0.547 |

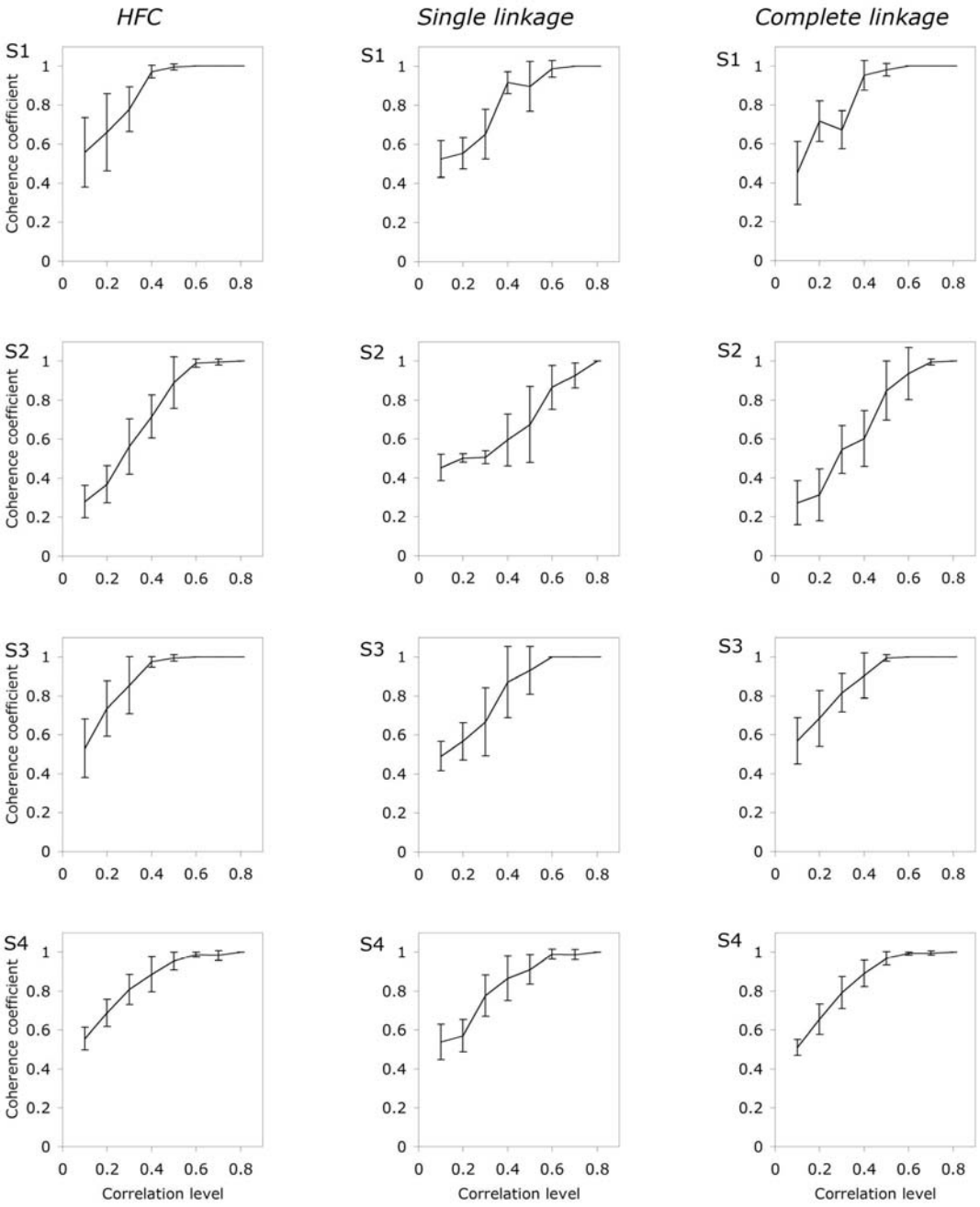


**Figure 2.** The four structures of the correlation matrices used in the experiments with different simulated data.

*S3*. The performances of complete and single linkage never differed significantly.

## Conclusions

*HFC* was introduced by Denimal (2001) to suggest a better procedure for the user than the classical one based on the tandem of PCA and hierarchical classification of units. The results of the experiments with simulated data indicated that *HFC* of variables performed as good as – and in some cases better than – both single and complete linkage clustering in detecting the known group structures. This certainly guarantees the consistency of its use, in comparison with these methods.

The experiments with simulated data confirm the consistency of the technique in detecting a predefined structure, at least at the same level of classical techniques used for classifying variables. So, the geometrical representations of both variables and units on principal planes, with their ability in getting easier the interpretation of results (see Camiz et al. 2006), is an extra feature of the results found, in respect to the others, without any loss in the intrinsic quality of the results.

**Acknowledgements:** The work described in this paper was developed by the authors during a stay of V. Pillar at Rome University La Sapienza as visiting professor and further reciprocal visits of the authors in their respective Universities. V. Pillar received also CNPq (Brazil) support. All institutions, involved in granting the visits, are gratefully acknowledged. Thanks are also due to J. J. Denimal who kindly provided the program for the *HFC* procedure and was generous in suggestions on the theoretical aspects of the method.

## References

Anderberg, M.R. 1973. *Cluster Analysis for Applications*. Academic Press, New York

Camiz, S., J.J. Denimal and V.D. Pillar. 2006. Hierarchical factor classification of quantitative variables and count data. *Community Ecology* 7: 165-179.

Denimal, J.J. 2001. *Hierarchical Factorial Analysis.* Proceedings of the 10th International Symposium on Applied Stochastic Models and Data Analysis. Compiègne, 12-15 Juin 2001.

Florek, K., J. Lukaszewicz, J. Perkal, H. Steinhaus and S. Zubrzycki. 1951. Sur la liason et la division des points d'un ensemble fini. *Colloquium Mathematicae* 2: 282-285.

Ganeshanandam, S. and W.J. Krzanowski. 1990. Error-rate estimation in two-group discriminant analysis using linear discriminant function. *Journal of Statistical Computation and Simulation* 36: 157-175.

Gordon, A.D. 1999. *Classification.* 2nd ed. Chapman and Hall, London.

Lance, G.N. and W.T. Williams. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer J.* 9: 373-380.

Legendre, P. and L. Legendre. 1998. *Numerical Ecology*, 2nd English edition. Elsevier, Amsterdam.

Lerman, I.C. 1991. Foundations of the likelihood linkage analysis (*LLA*) classification method. *Applied Stochastic Models and Data Analysis* 7: 63-76.

Milligan, G.W. and M.C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159-179.

Orlóci, L. 1991. *Entropy and Information.* SPB Academic Publishing, The Hague.

Peres-Neto, P.R. and D.A. Jackson. 2001. How well do multivariate datasets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* 129: 169-178.

Pillar, V.D. 1999. The bootstrapped ordination re-examined. *J. Veg. Sci.* 10: 895-902.

Pillar, V.D. 2006. *MULTIV: Multivariate Exploratory Analysis, Randomization Testing and Bootstrap Resampling, User's Guide v. 2.4.* Universidade Federal do Rio Grande do Sul, Porto Alegre.

Pillar, V.D. and L. Orlóci. 1996. On randomization testing in vegetation science: multifactor comparisons of relevé groups. *J. Veg. Sci.* 7: 585-592.

Podani, J. 2000. *Introduction to the Exploration of Multivariate Biological Data.* Backhuys, Leiden.

SAS Institute. 1999. *SAS Online Doc, Version 8.* SAS Institute Inc, Cary, North Carolina.

Sneath, P.H.A. 1957. The application of computers to taxonomy. *J. Gen. Microbiol.* 17: 201-226.

Sørensen, T. 1948. A method for establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter* 5(4): 1-34.

Vigneau, E., E.M. Qannari, K. Sahmer and D. Ladiray. 2006. Classification de variables autour de composantes latentes. *Rev. Statistique Appliquée* 54(1): 27-45.

Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc*. 58: 236-244.

Received July 7, 2006
Revised December 12, 2006
Accepted February 15, 2007