LEASP 97: AN IMPROVEMENT IN TEACHING AND ANALYSING NEW METHODOLOGY ON PROBABILISTIC CLUSTERING MODELS

Helena Bacelar-Nicolau*, F. C. Nicolau**, O. Dias*, Luis Ramos**, Portugal
* University of Lisbon(1); CEAUL/JNICT
** New University of Lisbon; CMAUNL/JNICT

Le logiciel LEASP 97 est la nouvelle version étendue portugaise du Logiciel Français LEAS d'Enseignement et Analyse Statistique, auquel ont été ajoutés notamment des programmes de classification sur des variables (approches classique et probabiliste) aussi bien que d'analyse des tableaux de contingence (classique et loglinéaire). Ses premières versions ont été developpés dans le cadre du programme de coopération scientifique et technologique luso-français ADAM1 d'Analyse des Données Multivariées, entre les laboratoires LEAD / FPCE / Université de Lisbonne, LEMA / Dep. Math. / Université Nouv. de Lisbonne et UB / Université de Montpellier. Dans ce travail nous analysons quelques aspects méthodologiques (associés aux modèles probabilistes) aussi bien que pédagogiques du LEASP 97, sur un ensemble de données issues d'un questionnaire concernant l'évaluation curriculaire d'un cours de l'enseignement supérieur universitaire.

LEAS 97

LEASP 97 is the new extended, improved Portuguese version of the French package LEAS for Teaching and Analysing Statistical Methodology. Its previous version (Bacelar-Nicolau, Nicolau, Mira, Dias,1994), has been developed into the Luso-French Scientific and Technological Program ADAM1(2) on Multivariate Data Analysis. The LEASP 97 is available in DOS.

LEASP 97 presents Portuguese menus and includes new subroutines, which have been developed by the Portuguese teams. Probabilistic hierarchical clustering models, which are now available in LEASP97, deal with classification of variables, a quite important problem in exploratory data analysis (very common in human sciences, for instance). Special attention has been paid to the case of multivariate discrete data analysis: hierarchical clustering algorithms for binary and frequency tables as well as classical and log linear algorithms for contingency tables, were implemented in order to combine exploratory and confirmatory methodology, in an easy manner, in the present version of LEASP. This allows us to use a general probabilistic approach for classifying multidimensional data.

The software LEASP 97 is organised in the same pedagogical way as the former LEAS: the tree of menus unfolds into chapters, sections and paragraphs as in a usual

course and/or manual of statistics and data analysis. It may be used as support for a course of three to four semesters. The hierarchy of menus is specially prepared for training the observation skills, critical capacity and accuracy of the students from collecting data and managing files to using new multivariate statistical methodology.

Here we shall analyse the LEASP 97 package in what concerns some methodological aspects (referred to the new algorithms) as well as pedagogical ones.

THE PROBABILISTIC APPROACH

Cluster analysis or classification names a set of multivariate methods for grouping elements (subjects or variables) from some finite set into clusters of similar elements (subjects or variables). Here we are mostly concerned with hierarchical clustering agglomerative models of variables (Bacelar-Nicolau, 1981, 1987, 1988; Gower, 1988; Lerman, 1981; Nicolau and Bacelar-Nicolau, 1998).

Classifying variables, that is partitioning a set of variables into classes or hierarchical groups of classes, is a major problem one faces in Exploratory Data Analysis, when dealing with data issued from human and social sciences and related areas. Particular attention has been paid by some researchers in this matter to the development of probabilistic clustering analysis techniques and methods which are able to deal with such kind of data (Bacelar-Nicolau, 1981, 1992; Lerman, 1981; Nicolau and Bacelar-Nicolau, 1998).

A classical hierarchical model can be simply represented by a bidimensional functional vector (*s* ,w) where *s* is a similarity (or dissimilarity) coefficient between variables and *w* an aggregation criterion between clusters of variables (Bacelar-Nicolau, 1981, 1987; Bacelar-Nicolau and Nicolau, 1994).

A probabilistic hierarchical agglomerative model can be defined as a four-dimensional vector (c, C; L, SD), where c is the (exact or asymptotic) cumulative distribution function (cdf) of a similarity coefficient (random variable) *S* on *s*, C being the (exact or asymptotic) cdf of an appropriate statistics W of the sample coefficient values, under some suitable null hypothesis (Bacelar-Nicolau, 1987, 1988, 1992; Lerman, 1981; Tiago de Oliveira, 1982); L is a probabilistic hierarchical levels index (Bacelar-Nicolau, 1981; Lerman, 1981) for selecting the most relevant levels in a hierarchy (from their absolute and local increasing maxima); and SD is another hierarchical levels index (Bacelar-Nicolau, 1981) for measuring the distortion degree of similarities in each level-

partition and/or detecting similarity conserving (stratification) groups of levels / partitions. The more general situation arises when c and C represent one or both some parametric / adaptive family of coefficients, instead of single ones (Nicolau, 1981, 1983; Nicolau and Bacelar-Nicolau, 1998). C might for instance represent an extension of the well known Lance and Williams adaptive formula for hierarchical agglomerative criteria. In that case by varying the parameters in a systematic way, robustness and/or validity studies based on the L and SD values can be accomplished inside such family, helping us to search for the models better fitting the data.

USING LEAS 97

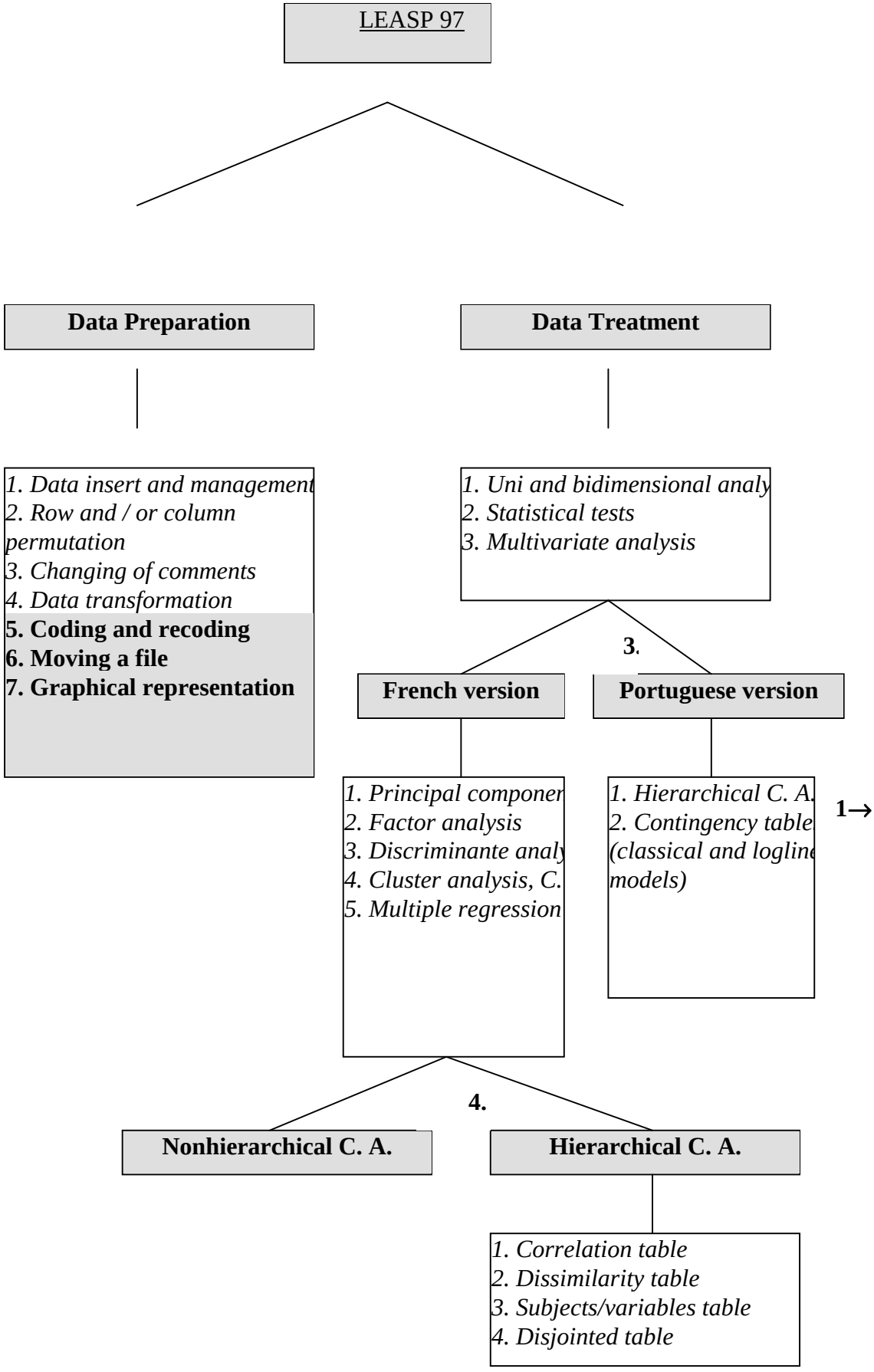
The next two pages present the general software scheme: the first one shows the main options allowed by the program; the second one shows the main options offered by the Portuguese version. A small manual is available providing basic instructions for using the software and the methodology. Note that in several cases a misinterpretation implies you to restart from the beginning, since we are concerned with a software for teaching and learning the statistical methods.

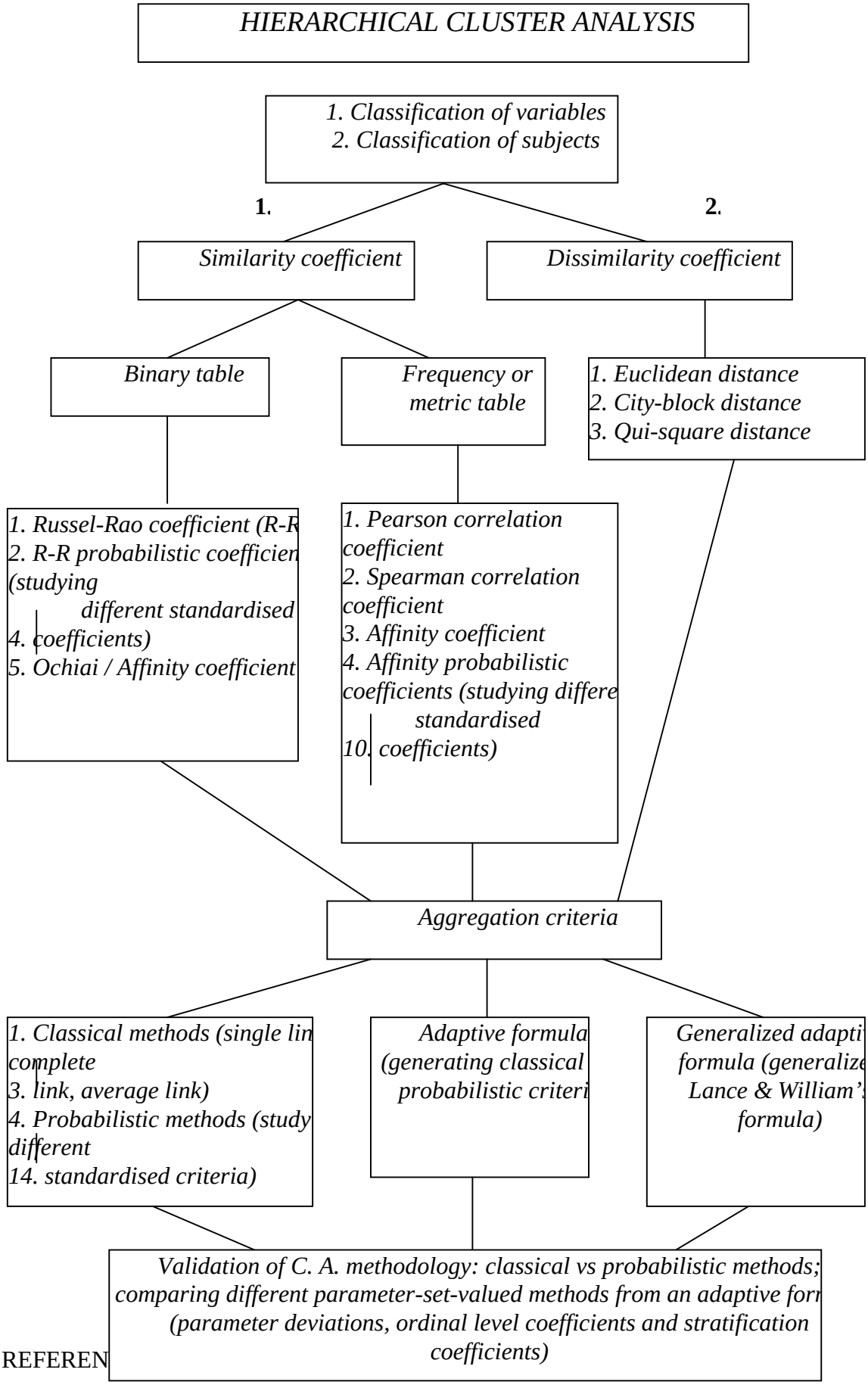
The Hierarchical Cluster Analysis scheme emphasises the separation between classification of observations which is mostly concerned with classical clustering methods and classification of variables mostly concerned with probabilistic methods, in the Portuguese module of LEASP 97. On the other hand one uses similarity coefficients in classification of variables. This can be used to learn and discuss about the role and the correspondence among similarity and dissimilarity coefficients in cluster analysis. One uses here only two *s* different measures of similarity in binary case, that are Russel-Rao and Ochiai coefficients. This allows us to discuss on the notions of monotone and distribution equivalence of coefficients, when we define the random *S* coefficient. The affinity coefficient (Bacelar-Nicolau, 1985, 1988; Matusita, 1955) is a similarity coefficient which we have studied either on the classical or the probabilistic point of view. We can remember here some important limit theorems of classical inference statistics (Tiago de Oliveira, 1982), concerning correlation coefficients, which allow us to use a probabilistic approach in cluster analysis and the (c, C; L, SD) model. Moreover we can relate affinity coefficient with other well known coefficients, like Pearson and Spearman coefficients, commonly used in clustering of variables. Finally we can point out the fact that affinity coefficient can deal with different type of data, since it generalises

the Ochiai coefficient for binary table and can be adapted to frequency as well as metric tables. In what concerns aggregation criteria the Portuguese module of LEASP 97 offers a set of probabilistic methods, either separately or included in some adaptive formulae. Also there are some classical aggregation criteria, separately or included in the adaptive formulae. This allows students to perform validation studies on the clustering results, specially when using the well known Lance and Williams adaptive formula for hierarchical agglomerative criteria. The validation studies particularly stand on comparison of clustering results obtained from suitable parameters variation and are based on computations of the corresponding L ordinal hierarchical level and SD stratification/distortion coefficients.

(1)* Laboratorio de Estatistica e Analise de Dados (LEAD), Faculdade de Psicologia e de Ciencias da Educação, Universidade de Lisboa. Alameda da Universidade, 1600 Lisboa, Portugal; Tel: 351.1.7934554; Fax: 351.1.7933408; Email: hbacelar@fc.ul.pt

(2)ADAM1: Scientific Research and Technological Luso-French Cooperation Program on Multivariate Data Analysis, supported by the JNICT/Portugal and the Embassy of France, joining the Laboratory of Statistics and Data Analysis (LEAD) at the Faculty of Psychology and Education, University of Lisbon / Prof. H. Bacelar-Nicolau, the Laboratory of Statistics and Actuarial Mathematics (LEMA) / Department of Mathematics / New University of Lisbon / Prof. F.C. Nicolau and the Biometry Unity, INRA, University of Montpellier / Prof. Y. Escoufier.


LEASP 97
Data Preparation
Data Treatment
1. Data insert and management
2. Row and / or column permutation
3. Changing of comments
4. Data transformation
5. Coding and recoding
6. Moving a file
7. Graphical representation
1. Uni and bidimensional analy
2. Statistical tests
3. Multivariate analysis
3.
French version
Portuguese version
1. Principal componer
2. Factor analysis
3. Discriminante analy
4. Cluster analysis, C.
5. Multiple regression
1. Hierarchical C. A.
2. Contingency table (classical and logline models)
1→
4.
Nonhierarchical C. A.
Hierarchical C. A.
1. Correlation table
2. Dissimilarity table
3. Subjects/variables table
4. Disjointed table


HIERARCHICAL CLUSTER ANALYSIS
1. Classification of variables
2. Classification of subjects
1.
2.
Similarity coefficient
Dissimilarity coefficient
Binary table
Frequency or metric table
1. Euclidean distance
2. City-block distance
3. Qui-square distance
1. Russel-Rao coefficient (R-R
2. R-R probabilistic coefficien
(studying
different standardised
4. coefficients)
5. Ochiai / Affinity coefficient
1. Pearson correlation coefficient
2. Spearman correlation coefficient
3. Affinity coefficient
4. Affinity probabilistic coefficients (studying differe
standardised
10. coefficients)
Aggregation criteria
1. Classical methods (single lin
complete
3. link, average link)
4. Probabilistic methods (study
different
14. standardised criteria)
Adaptive formula (generating classical probabilistic criteri
Generalized adapti formula (generalize Lance & William's formula)
Validation of C. A. methodology: classical vs probabilistic methods; comparing different parameter-set-valued methods from an adaptive forr (parameter deviations, ordinal level coefficients and stratification coefficients)

REFEREN

Bacelar-Nicolau, H.(1981). *Contributions to the Study of Comparison Coefficients in Cluster Analysis*, Univ. Lisbon.

Bacelar-Nicolau, H.(1985). The affinity coefficient in cluster analysis, *Meth. Oper. Res., 53,* Verlag Anton Hain,. 507-512.

Bacelar-Nicolau, H. (1987). *On the distribution equivalence in cluster analysis*, NATO ASI Series, vol. F30, Pattern Recognition Theory and Applications, P.A. Devijver/J. Kitler (eds.), Springer-Verlag.

Bacelar-Nicolau, H.(1988). *Two probabilistic models for classification of variables in frequency tables*, Classif. and Relat. Meth. of Data Analysis, H. .H. Bock (ed.), North Holland, 181-186.

Bacelar-Nicolau, H.(1992). *Probabilistic similarity coefficients for variables in hierarchical agglomerative clustering models*, Abstracts of the International Meeting on Distance Analysis, DISTANCIA'92, 93-94.

Bacelar-Nicolau, H., C. Nicolau, F. (1994). *Exploratory and confirmatory discrete multivariate analysis in a probabilistic approach for studying the regional distribution of AIDS in Angola*, New Approaches in Classification and Data Analysis, E. Diday /Y. Lechevalier/M. Schader (Eds.), Springer-Verlag, 610-618.

Bacelar-Nicolau, H., C. Nicolau, F., Mira, C., Dias, O., (1994). *LEASP: Learning and Teaching New Methodology on Probabilistic Clustering Models.* Contributed paper presented at the International Conference on Teaching Statistics (ICOTS4), Marrakech, Morocco.

Gower J. C.(1988). *Classification, geometry and data analysis.* Classification and Related Methods of Data Analysis, H. H. Bock (Ed.), North Holland, 3-14.

Lerman I. C.(1981). *Classification et Analyse ordinale des Données*, Dunod, Paris.

Matusita K. (1955). *Decision rules, based on distance for problems of fit, two samples and estimation*, Ann. Inst. Stat. Math., 26, n.º4, pp. 631-640.

Nicolau, F.C.(1981). *Hierarchical Cluster Analysis Criteria Based on Distribution Function.* PhD Th., Univ. Lisbon.

Nicolau F.C. (1983). *Cluster analysis and distribution function, Meth. Oper. Res., 45,* Verlag Anton Hain, 431-433.

Nicolau F.C., Bacelar-Nicolau, H. (1998). *Some Trends in the Classification of variables*, Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, Y. Baba (Eds.), Spriger, 89-98.

Tiago de Oliveira, J.(1982). *The Delta-Method for Obtention of Asymptotic Distributions: Applications*. Publ. Inst. Stat. Univ. Paris, vol. XXVII, 49-70.