Acta Scientific COMPUTER SCIENCES

Volume 4 Issue 10 October 2023

Research Article

Comparison between Simple Correspondence Analysis and Canonical Correspondence Analysis: Application in Public Health

Estefania J Guevara[1], Melba L Vertel[2]* and Daniel E Tamara[1]

[1]Department of Mathematics and Statistics, Faculty of Exact and Natural Sciences, National University of Colombia, Manizales, Colombia

[2]Department of Mathematics, Faculty of Education and Sciences, University of Sucre, Sincelejo, Colombia

***Corresponding Author:** Melba L Vertel, Department of Mathematics, Faculty of Education and Sciences, University of Sucre, Sincelejo, Colombia.

Received: July 29, 2023

Published: September 28, 2023

Abstract

The main objective of this work is to methodologically compare simple correspondence analysis (ACS) and canonical correspondence analysis (ACC) applied to frequency tables. A theoretical presentation of the weighted principal component analysis (PCA) is made under the "French school of data". The comparison of the two methods refers to putting them in parallel, since they do not point to exactly the same methodological objectives; properties, common and different elements of the methods are presented and illustrated with the example of Urbina and Londoño (2003). This analysis methodology is presented with an application in public health based on the work of Iriarte., et al. (2012). To execute the statistical techniques, the R software, the ade4 and FactoClass packages are used.

Keywords: Multivariate Data Analysis; Simple Correspondence Analysis; Canonical Analysis of Correspondences; Public Health; Statistical Language R

Introduction

The application of Statistics is today a fundamental tool in the analysis, conclusions and recommendations of scientific research. It is common that in statistical analysis only the methods of univariate descriptive statistics (analysis and representation of data in numerical and graphic form) are used, conforming the researcher (apparently) with the simple description through uní or two-dimensional analysis, although it is necessary to do so because it allows a first approach to the characteristics of the information, many times, as has been seen in different publications, they are not always the most appropriate for the solution of the proposed problems, nor to achieve the objectives set out in these investigations.

Univariate and bivariate descriptive methods are part of the courses and texts of statistics or quantitative methods of the academic programs of the different disciplines [1-7], although they are the multivariate methods of data [8-18], which make it possible to take into account the interrelationship between multiple variables.

The nature of the rows and columns of a data table together with the objectives of the study determine the statistical methods to be used. Although the information collected is multivariate in nature and the progressive and sustained evolution of computer science in the last 20 years allows multivariate techniques to be increasingly used and the management of software and computational resources to automate tasks of elaboration, analysis of results and the design and implementation of relational and multidimensional databases are more friendly, this type of analysis is practically non-existent in many research papers.

The object of study in this work is the analysis of frequency tables, the result of the growing volume of presence-absence data, counts or percentages of economic, social, natural, psychological, geographical, historical or political phenomena to extract knowledge and serve as support for decision making, which becomes a problem and an opportunity that requires the definition and implementation of data analysis and pattern recognition techniques.

To analyze frequency tables (contingency tables, absence-presence, counts, percentages), the most useful multivariate descriptive method in applied sciences is simple correspondence analysis (DHW) [16,19-30].

The ACS seeks the best simultaneous representation of two sets, consisting of the rows and columns of a frequency table, through a reduction in dimension that allows the noise to be isolated to examine the relationships between the variables [31-34].

ACS can be seen as the simultaneous application of two principal component analyses (PCAs). In the ACS additional variables can be used to analyze pre-established objectives, just as in an ACP on the factorial axes you can project rows and columns that have not participated in the analysis.

It may happen that we want the result of the ACS to be related to external variables, which have an active role in the definition of the results of the frequency table. The multivariate technique that helps us do this is the Canonical Correspondence Analysis (ACC) proposed by Ter-Braak [35], frequently used in ecology to study the influence of environmental conditions on the distribution of species of flora and fauna [14,36-41]. In these cases, when carrying out the ACS we would look for the sub-spaces that best explain the data in the frequency table, but with the condition that these are related to the external variables (quantitative or qualitative).

The main objective of this work is to methodologically compare the ACS and the ACC applied to frequency tables [38]. In addition, a theoretical presentation of the analysis in weighted principal components (PCA) is made under the "French school of data" to find meaning to the information collected through projections on lines and planes [37].

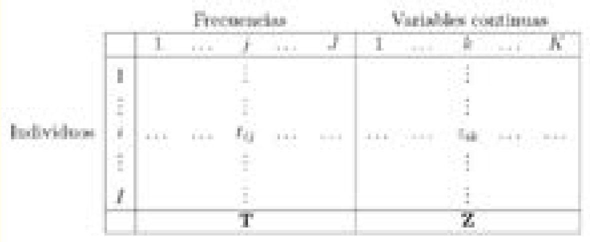
The comparison of the two methods refers to putting them in parallel, since they do not point to exactly the same methodological objectives; properties, common and different elements of the methods will be presented and illustrated with the example of Urbina and Londoño [41].

In order to contribute to the dissemination of multivariate statistical techniques, the ACS and the ACC applied in an area other than environmental research will be analyzed: public health.

Multivariate descriptive techniques

Problems arising from the social sciences, biology, and health sciences generally require the use of multivariate descriptive techniques to present the original data in a summarized manner and facilitate understanding of them through graphical representations.

The databases that are considered have the form [TZ], where T is a table of frequencies whose inputs are expressed in absolute terms or in percentage and Z is a table of continuous variables of quantitative data. The table [T Z], illustrated in figure 1, is called the table of frequencies - continuous variables.



		Frecuencias					Variables continuas				
		1	…	j	…	J	1	…	k	…	K
	1			⋮					⋮		
	⋮			⋮					⋮		
Individuos	i	…	…	t_{ij}	…	…	…	…	z_{ik}	…	…
	⋮			⋮					⋮		
	I			⋮					⋮		
				T					Z		

Figure 1: Table of frequencies - continuous variables.

$T = (t_{ij})$of order and is of order. We define as the table of relative frequencies corresponding to T, that is,$I \times JZ = (z_{ik}) I \times KF = (f_{ij}) f_{ij} = \frac{t_{ij}}{s}$, where. The values and $s = \sum_{ij} t_{ij}\, f_{i.} = \sum_{j=1}^{J} f_{ij} = \sum_{j=1}^{J} \frac{t_{ij}}{s} f_{.j} = \sum_{i=1}^{I} f_{ij} = \sum_{i=1}^{I} \frac{t_{ij}}{s}$ are called, respectively, marginal row and marginal column of the table. The diagonal matrices and. $FD_I = diag(f_{i.}) D_J = diag(f_{.j})$

Principal Component Analysis (PCA)

In the GPA it is considered a matrix or table of initial data of a non-symmetric nature. Geometrically, the ACP transforms this data table into two point clouds: one of individuals and one of variables. In the first, individuals are compared and in the other, the relationships between the variables are studied. To interpret point clouds it is necessary to project them on lines and planes so that as much information as possible is preserved. The similarity between individuals is determined by a geometric distance (the resemblance or difference of individuals is translated into their proximity or remoteness) and the correlation between variables by the angle formed by the vectors that represent them.

On the other hand, within the objectives of the ACP are: 1. Compare individuals from the continuous variables, 2. Identify the relationships between the variables and 3. Reduce the dimensionality of the problem.

Weighted principal component analysis

The weighted PCA is an ACP of an X matrix that contains the data to be analyzed (standardized). The weighted ACP is denoted by *ACP (X, M, D),* where and are, respectively, the diagonal matrices of the weights of the columns and rows. *MD*

Cloud		
Space		
Metric	M	D
Coordinates	Rows of X	Columns of X
Weight	Diagonal of D	Diagonal of M
Inertia	Trace (X'DXM)	Trace (XDX'M)
Eigenvalue		
Eigenvector		
Factorial coordinates		
Transition formulas		

Table 1: Formulas of the ACP (X, M, D).

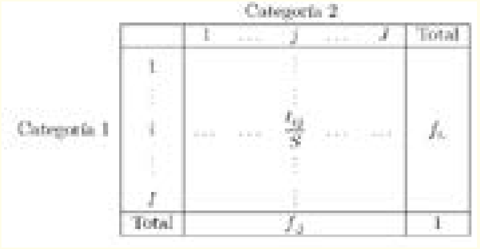
The main formulas of the are shown in table 1 *ACP (X, M, D)*.

Simple correspondence analysis (ACS)

Contingency tables, common in the social sciences, are used to organize and analyze the relationship between two or more qualitative variables. Simple correspondence analysis is a statistical technique used in the analysis, from the geometric point of view, of the relationships of a set of frequency variables (counts, binary responses or percentages) organized in a contingency table.

The ACS associates to each of the categories of the contingency table a point in the 2-dimensional space, so that the proximity or remoteness between these points means dependence or similarity between them.

The ACS of the frequency table is the, with (the matrix of ones, of dimension) and, defined as before. $TACP(P,D_J, D_I)P=D_I^{-1} FD_J^{-1}-1_{IJ}$ $1_{IJ}\ I\times JD_I\ D_J$



		Categoria 2					
		1	…	j	…	J	Total
	1						
Categoria 1	i	…	…	$\frac{f_{ij}}{S}$	…	…	$f_{i.}$
	I						
	Total			$f_{.j}$			1

Figure 2: Table of relative frequencies.

The table of relative frequencies (sometimes called table or correspondence matrix), of, is shown in figure 2.
The elements shown in figure 2 refer to those defined in section 1.

Remarks

The general term of is given by. ${}^{P}p_{ij} = \frac{f_{ij}-f_{i.}f_{.j}}{f_{i.}f_{.j}}$.
The I row in the table make up the point cloud at. $TN_J^I\mathbb{R}^J$
The J columns in table T make up the point cloud at. $N_J\mathbb{R}^I$

Canonical Correspondence Analysis (ACC)

The ACC (35) is a method that allows to simultaneously analyze a group of frequencies and a group of variables on the same set of individuals. The ACC only takes into account the part of the structure of the frequency table that can be explained by the continuous variables.

In the ACC of the table (see Figure 1), it is taken as a set of response or dependent variables and as another of independent or explanatory variables. $[T\ Z]TZ.$

The ACC of the table of frequencies - continuous variables is the, with the standardized table of and as before. $[T\ Z]\ ACP(\hat{Y}, D_J, D_I)\hat{Y} = P_{Z_0}D_I^{-1}FD_J^{-1},\ P_{Z_0} = Z_0(Z_0'D_IZ_0)^{-1}Z_0'D_I, Z_0 Z D_J, D_J.$

Methodological guide for the practical implementation of the ACS and the ACC

The ACS and ACC methods act on databases that contain a portion of counting variables and a portion of continuous variables (Tables 3 and 5). They analyze the relationship between such variables to identify which ones have the greatest influence on a given situation.

Below is a methodological guide to apply each of these methods.

Simple Correspondence Analysis (ACS)

Creation of a database of the form, where is a table of frequencies whose inputs are expressed in absolute terms or in percentages and is a table of continuous variables of quantitative data.. $[T\ Z]TZ.$

Performing a correspondence analysis of the frequency matrix.

Performing an analysis on main components of the group of variables continues to take into account the correspondence analysis performed in step 2.*Z.*

Elaboration of factorial planes and circle of correlations.

Interpretation of the graphs resulting from 4.

Canonical Correspondence Analysis (ACC)

Creation of a database of the form where it is a table of frequencies whose inputs are expressed in absolute terms or in percentages and is a table of continuous variables of quantitative data.. $[T\ Z]TZ.$

Performing the canonical analysis of correspondences of the table of frequencies - continuous variables..

Elaboration of the factorial plane and circle of correlations.

Interpretation of the graphs resulting from 4.

Comparison between ACS and ACC methods

In this section a methodological comparison of the ACS and ACC methods is made, where some theoretical characteristics of them are put in parallel (see Table 2).

Method	ACS	ACC
Cloud of individuals		
Space of individuals		
Variable Cloud		
Variable space		
Weight of individuals		
Matrix X		
Weighting of variables		
ACP (X, M, D)		
Inertia		
Eigenvalue		
Transition formula for rows		
Transition formula for columns		

Table 2: Theoretical comparison between ACS and ACC methods.

Common aspects

The TISA and the PCA are weighted ACP, wherein according to section 1.1.1 the matrices to be analyzed are, respectively, and. Specifically, $X = PX = \tilde{P}ACS(T) : ACP(P, D_J, D_I)$ and. $ACC(T, Z): ACP(\tilde{P}, D_J, D_I)$ As an additional data you have to. $ACP(Z): ACP(Z_0, D_I, D_I).$

ACS and ACC have in common the diagonal matrix of weights of the columns and the diagonal matrix of weights of the rows. .

$M = D_J D = D_I.$
The study of associations between individuals and frequencies.

Different aspects

The ACS only analyzes associations between individuals and frequencies; the ACC, on the other hand, also studies the dependency relationships that such frequencies have with the external group of continuous variables.

The inertia of the always will take a value less than or equal to the inertia of the ACS of the frequency table. *ACC (T, Z).*

Comparison between ACS and ACC for the GORGONA example

The work of Urbina and Londoño [41] will be used to illustrate and compare the ACS and ACC methods. This study sought to know the distribution of the herpetofauna community on the island of Gorgona and determine the possible relationship of some species with temperature, relative humidity and vegetation cover on microhabitats.

The information collected in Urbina and Londoño [41] is shown in table 3.

The absolute frequency table crosses 32 rows (sections located in the different areas of Gorgon Island) and 11 columns (reptile and amphibian species). The table of continuous variables crosses the same rows and 5 columns (variables related to climate and habitat). *TZ.*

Analysis of the Gorgon example with ACS

The total inertia associated with the ACS of the frequency table is 1,308 and is represented by the first two eigenvalues by 75.2%. It is observed from the factorial plane of (Figure 3) that the first axis separates the species R. venenosa (present in the areas of prison and crops) from the species R. Brincona and R. Arlequin (present in the areas of primary and secondary forests). *TACS(T).*

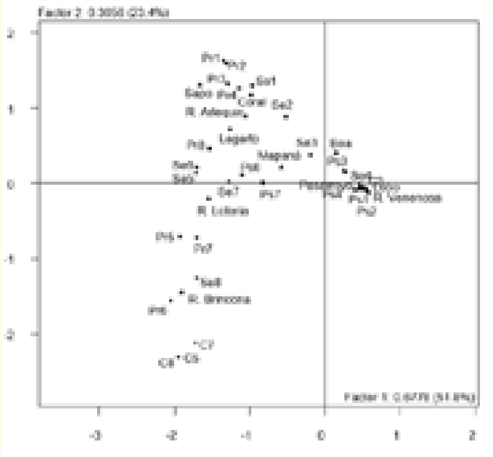


Figure 3: Plano factorial del ACS (T).

For the, that is, of the climate and habitat variables, the inertia is 5 and is represented by 69.1% by the first two eigenvalues (Figure 4). The variables that contribute most in the representation are temperature, arbústica coverage, herbaceous cover and canopy cover. The first axis contrasts areas with high levels of shrub and canopy cover with low index areas in such coverages. Axis 2 separates sections of high temperature indices and herbaceous cover with those of low temperature level and herbaceous cover. *ACP(Z)*

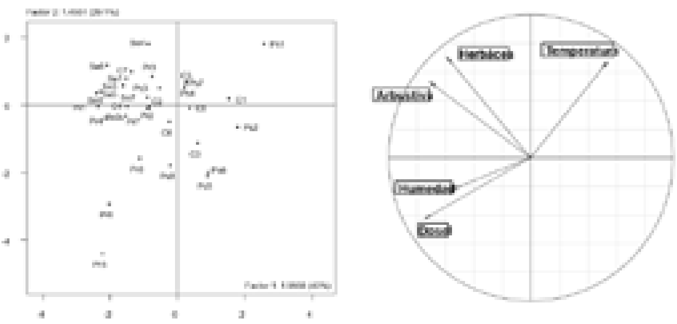


Figure 4: Factorial plane and circle of correlations of the ACP (Z).

Analysis of the GORGONA example with ACC

The inertia associated with the is 0.3948 (Figure 5) and is represented by 91.8% by the first two components. There is therefore a good representation of the relationship between climate variables - habitat and species. 81.1% of the accumulated inertia on the first axis indicates that the continuous variables (temperature, humidity and arbústic, herbaceous and canopy coverages) satisfactorily explain this factor. *ACC(T, Z).*

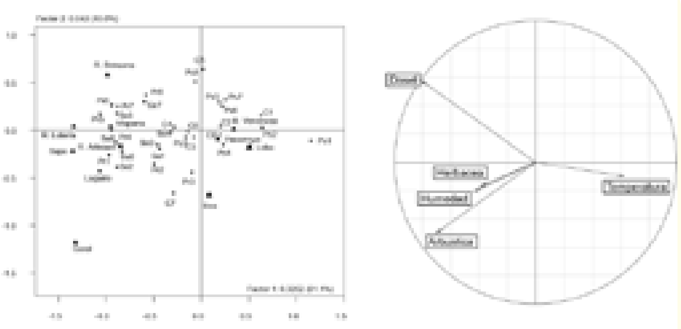


Figure 5: Factorial plane and circle of correlations of the ACC (T, Z).

ACS vs ACC: Gorgon Example

The factorial plane of the ACC represents 91.8% of the variables and the ACS 75.2%. Factor 1 in the ACC explains 81.1% () and in the ACS 51.8% () of the total variability, meaning that the ACC makes a better representation of the original variables.

Comparison between ACS and ACC in frequency tables: application in public health

The illustration and comparison of the ACS and ACC methods in the area of public health is carried out through the work of Iriarte., *et al*. (2012). This research sought to detect from a sample of 226 rodents (Mus Musculus, Ratus Norvegicus and R. Rattus) which had the bacterium Leptospira spp causing Leptospirosis (disease transmitted from animals to humans).

The information collected in Iriarte., *et al*. (2012) is shown in table 4.

The frequency table crosses 10 rows corresponding to areas of the municipality of Sincelejo with characteristics of rodents (species, sex, maturities, number of infected). The board crosses the same rows with some physical features of such animals (weight, total length, tail, ear). *T Z.*

ACS Application in Public Health

In the ACS of the frequency table, a total inertia of 0.0753 was obtained, represented by 77.1% by the first two own values. From the factorial plane of (Figure 6) it is noted that the second axis

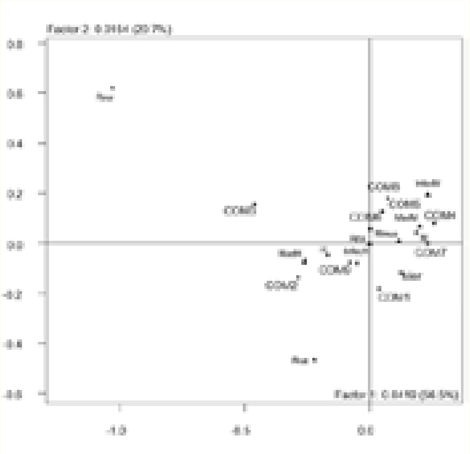


Figure 6: Plano factorial del ACS (T).

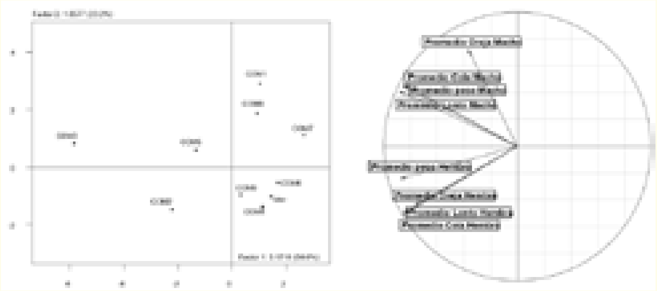


Figure 7: Factorial plane and circle of correlations of the ACP (Z).

separates the two least found species, Rattus Norvegicus (present mostly in commune 3) and R. Rattus (present mainly in commune 2 and market). *TACS(T).*

As for the, the first two eigenvalues represent 87.9% of the total inertia, this being 8 (Figure 7). It is highlighted that all variables, with the exception of average male ear, have a great contribution in the representation of the data in the factorial plane; communes 2, 3 and 5 have a good quality of representation. *ACP(Z).*

ACC Application in public health

The total inertia associated with the is 0.0611 and is supported by 81.4% in the first two axes, therefore there is a good representation of the relationship between the study variables. It is important to note that the communes where there is a greater amount of the species Ratus Norvegicus, variable of greater contribution, are the 2 and 3, same where the continuous variables have high percentages. *ACC (T, Z).*

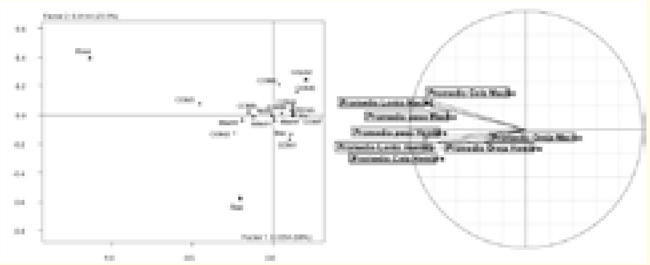


Figure 8: Factorial plane and circle of correlations of the ACC (T, Z).

ACS vs ACC: Application in Public Health

The factorial plane of the ACC represents 81.4% of the variables and the ACS 77.1%, that is, the ACC represents the original variables better.

Conclusion

As a result of applying simple correspondence analysis (ACS) and canonical correspondence analysis (ACC) to ecological (e.g., Gorgona) and public health (leptospira) databases, it is concluded that the second method better represents the original variables within a factorial plane. Likewise, by describing the dependence between frequencies and continuous variables, the ACC allows a better interpretation of the data.

Thanks

This work has been carried out thanks to an agreement between the University of Sucre and Colciencias in which it seeks to train young researchers and innovative professionals, through the granting of an internship grant in the traditional way. We also thank the work carried out by members of the research group “biomedical research” (42-43-44) of the University of Sucre.

To execute the statistical techniques, the R software [45] and the ade4 [12] and Facto Class [16] packages are used.

Bibliography

1. Mendenhall W. “Statistics for administration and economics, Vol. III, I beroamerica, Mexico City” (1981).

2. Spiegel M and D Murray. “Statistics, Vol. VII, Mc Graw Hill, Mexico City” (1991).

3. Canavos G. “Probability and Statistics: Applications and Methods, Mc. Graw Hill” (1992).

4. Walpole R and R Myers. “Probability and Statistics, Mc Graw Hill” (1997).

5. Montgomery D and Runger C. “Probability and Statistics applied to Engineering, Mc. Graw Hill” (1996).

6. Weimer P. “Statistics, CECSA” (1999).

7. Martínez C. “Estadística y Muestreo, segunda edn, ECOE, Universidad Nacional de Colombia, Bogotá” (2004).

8. Lebart L., *et al*. “Statisitique exploratoire multidimension, Dunod, Paris” (1995).

9. Lebart L., *et al*. “Statistical data processing, Barcelona, Marcombo” (2000).

10. Escofier B and J Pagès. “Single and multiple factor analysis. Objectives, methods and interpretation, University of the Basque Country, Bilbao” (1988-1998).

11. Cabarcas G and C Pardo. “Multivariate statistical methods in social research, Cursillo, Statistics Symposium - Santa Marta. National University”. *Department of Statistics* (2001).

12. Dray SyD. “Elèments d’interface entre analyses multivariées, systèmes d’information geographique et observations écologiques, PhD thesis, Universite Claude Bernard - Lyon 1 (2003).

13. Chessel DyA and Dufour. “The ade4-package: implementing the duality diagram for ecologists”. *Journal of Statistical Software* 22 (2007): 1-20.

14. Greenacre M. “Correspondence Analysis in Practice, 2nd edition, Chapman and Hall/CRC (2007).

15. Pardo CE. “Euclidean geometry in statistics: methods in main axes, Department of Statistics, National University of Colombia, Bogotá”. Conference for the V Conference on Mathematics, Statistics and Didactics. Pedagogical and Technological University of Colombia (2008).

16. Pardo CE and PDel Campo. “Combination of factorial methods and cluster analysis in R: the FactoClass package”. *Revista Colombiana de Estadística* 30 (2007).

17. Díaz L and M Morales. “Statistical analysis of categorical data, master’s thesis, Faculty of Sciences, Department of Statistics, National University of Colombia (2010).

18. Husson F., *et al*. “Exploratory Multivariate Analysis by Example Using R, Chapman and Hall/CRC computer science y data analysis (2011).

19. Hirschfeld HO. “A connection between correlation and contingency, Proceedings of the Cambridge Philosophical Society”. *Mathematical and Physical Sciences* 31 (1935): 520-524.

20. Fisher R. “The precision of discriminant functions”. *Annals of Eugenics* 10 (1940): 422-429.

21. Guttman L. “The quantification of a class of attributes: A theory and method of scale construction”. *Social Science Research Council, New-York* (1941): 319-348.

22. Hayashi C. “On the quantification of qualitative data from the mathematic statistical point of view”. *Annals of the Institute of Statistical Mathematics* 2 (1950): 35-47.

23. Nishisato S. “Analysis of Categorical Data: Dual Scaling and its Applications, University of Toronto Press, London (1980).

24. Williams E. “Use of scores for the analysis of association in contingency tables”. *Biometrika* 39 (1952): 274-289.

25. Hill M. “Reciprocal averaging: an eigenvector method of ordination”. *Journal of Ecology* 61 (1973): 237-249.

26. Hill M. “Correspondence analysis: A neglected multivariate method. applied statistics”. *Journal of the Royal Statistical Society Series C* 23 (1974): 340-354.

27. Benzecri J. “Statistical analysis as a tool to make patterns emerge from data, dans Watanabe S., ed”. *Methodologies of Pattern Recognition* (1969): 35-60.

28. Cailliez F and yJ Pagès. “Introduction à l’analyse des données, SMASH, 9 rue Duban 75016 Paris”. (1976): 616.

29. Tenenhaus M and YF Young. "An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data". *Psychometrika* 50 (1985): 91-119.

30. Escoufier Y. "The duality diagram: a means of better practical applications, Legendre. and L. Legendre, editors". Development in numerical ecology, Springer Verlag, Berlin, Germany (1987): 139-156.

31. Fine J. "Introduction to multidimensional data analysis from examples, Brochure, PRESTA: Programme de recherche et a'enseignement en statistique appliquée, Sao Carlos (1996).

32. Fernández P. "The use of simple correspondence analysis (ACS) as an aid in the interpretation of data in archaeology. A case study". *Anthropological Bulletin, Universidad de Los Andes* 55 (2002): 687-713 (2002).

33. Becué M., *et al*. "Multiple factor analysis for contingency tables: study of mortality in the autonomous communities of Spain, National Congress of Statistics and Operations Research, Lleida 8 to 11 April (2003).

34. Vertel M and CE Pardo. "Comparison between canonical correspondence analysis and multiple factor analysis in tables of continuous frequencies-variables, Master's Thesis, National University of Colombia, Faculty of Sciences". *Department of Statistics, Bogotá* (2010).

35. Ter-Braak C. "Canonical correspondence analysis: A new technique for multivariate direct gradient analysis". *Ecology* (1986): 65.

36. Chessel D., *et al*. "Propriétés the canonical analysis of correspondences; an illustration in hydrobiology". *Revue Statistique Appliquée* 35.4 (1987): 55-72.

37. Lebreton J., *et al*. "Principal component and correspondence analyses with respect to instrumental variables: an overview of their role in studies of structure-activity and species-environment relationships". *Applied Multivariate Analysis in SAR and Environmental Studies* (1991): 85-114.

38. Dolédec S and YD Chessel. "Recent developments in linear ordination methods for environmental sciences". *Advances in Ecology, India* 1 (1991): 133-155.

39. Birks H and H Austin. "An annotated bibliography of canonical correspondence analysis and related constrained ordination methods, Botanical Institute, Norway". All-Gaten 41, N-5007 Bergen, Bunch, K.J., Heneghan (1994).

40. Pavoine S and YA Dufour. "Canonical correspondence analysis, a standard in ecology, CARME 2003: International Conference on Correspondence Analysis and Related Methods (2003): 63-64.

41. Urbina J and M Londoño. "Distribution of the herpeto-fauna community associated with four areas with different degrees of disturbance on Gorgona Island, Colombian Pacific". *Revista de la Academia Colombiana de Ciencias* 27 (2003): 105-112.

42. Iriarte I., *et al*. "Seroprevalencia a leptospira spp patógeno en rodentes del área urbana de la ciudad de Sincelejo, Sucre, Memorias: XXII Simposio de Estadística UNAL". Faculty of Sciences. Department of Statistics, Bucaramanga (2012).

43. Castellar A., *et al*. "Identification of leptospira interrogans in a population of rodents from two localities of the municipality of sincelejo, degree work, University of Sucre (2012).

44. Perez A., *et al*. "Identification of leptospira interrogans by polymerase chain reaction (pcr) in a population of rodents in the municipality of sincelejo, Degree work, University of Sucre" (2013).

45. R Development Core Team T: "A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2007).