Epidémiol. santé anim., 1997, 31-32

DECENTRED CORRESPONDENCE ANALYSIS : A METHOD FOR ANALYSING A MULTIVARIATE TABLE OF PROPORTIONS

Tillard E.[1], Faye B.[2], Desoutter D.[3], Chessel D.[4], Thioulouse J.[5]

Une enquête séro-épidémiologique longitudinale a été réalisée chez les ovins élevés dans la zone sahélienne au Sénégal, pour préciser l'étiologie infectieuse des affections respiratoires. Dans 9 villages, des prélèvements sanguins ont été réalisés à intervalle régulier sur des animaux, entre l'âge d'1 mois et de 15 mois. Pour chaque prélèvement, 3 virus (parainfluenza III, Adénovirus ovin type 5, virus respiratoire syntitial) et 9 types de pasteurelles (genres multocida et haemolytica) ont été recherchés. Les résultats sérologiques successifs déterminent une cinétique d'anticorps pour chacune des 4 saisons de la période d'étude. Dans le but d'identifier les associations et les compétitions entre agents pneumotropes au niveau villageois, et leurs variations spatio-temporelles, on a construit un tableau croisant en ligne les villages et les saisons, en colonne les agents infectieux, et contenant à l'intersection des 2 un « taux de circulation », défini comme le rapport du nombre d'animaux présentant une cinétique ascendante à la totalité des animaux prélevés.

On propose d'analyser un tel tableau de taux par l'Analyse Factorielle des Correspondances Décentrée (AFCD), extension de l'AFC simple, dans laquelle la marge des lignes est remplacée par une marge extérieure imposée contenant les effectifs d'animaux prélevés. Contrairement aux autres analyses envisageables (Analyse en Composantes Principales, AFC), l'AFCD donne aux profils une signification épidémiologique: le profil ligne représente le taux de circulation pour un village, une saison et agent donné, et le profil colonne, le taux de circulation moyen pour un agent donné (moyenne de variables quotients). Au total, l'analyse permet de comparer les villages entre eux et les périodes entre elles, même si les effectifs prélevés sont très différents.

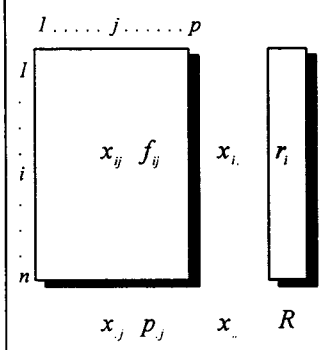
INTRODUCTION

Respiratory diseases often represent the main ruminant diseases encountered in most African sahelian countries. To identify the most important infectious agents in sheep in the northern area of Senegal, a longitudinal serological survey was carried out by the PPR programme : a research programme on small ruminants raised in traditionnal farm settings implemented by both CIRAD-EMVT[6] and ISRA-LNERV[7]. The principal infectious agents were then analyzed one by one and the data gattered permitted comparisons between zones, villages, and age spans (using the χ^2 test) (Desoutter, 1994). However, this univariate approach is too restrictive for studying a multifactorial pathology, involving at the same time association of several infectious agents and environmental factors linked to breeding practices. Multivariate analysis techniques then were undertaken to identify association or competition between lung infection agents observed at the village level, and to define their spatial and temporel variations, or, in other words, to look into the correlations between dependent variables and the fixed effects (villages, periods).

I. MATERIAL AND METHODS

The survey was carried out in 2 sahelian areas (Louga and St Louis) between november 1988 and july 1989, i.e. a space of time including successively 4 periods, a dry hot season (HS1), a rainy season (RS), a dry cold season (CS) where respiratory diseases may occur more frequently, and a second dry hot season (HS2). Blood was sampled in young animals of over one month up to 15 months old, every 4-6 weeks, in 9 villages with a sizeable population. Twelve lung infectious agents were systematically looked for in each blood sample using serological techniques : 3 virus, *Parainfluenza III* (PI3), *Respiratory Syncytial Virus* (RSV), A*denovirus type 5* (AD5) and 9 *Pasteurella serotypes* (*haemolytica 1, 2, 5, 6, 7, 9, 11,* written *Hx* and *multocida A* et *D* written *MA* and *MD*). For each period and village, an indicator of current infection is calculated by dividing the sum of animals presenting a rising titre of antibodies by the sum of animals sampled twice, both at the beginning and at the end of the study period. The ratio were then adjusted according to age at the entry into survey, for each period. Finally, a raw data table having n=26 rows (*i*), and p=12 columns (*j*) was analysed.

[1] CIRAD-EMVT, Ligne Paradis, 97410, St Pierre, La Réunion, France
[2] CIRAD-EMVT, Campus de Baillarguet, BP 5035, 34032, Montpellier cedex 1, France
[3] Nouméa, Nouvelle Calédonie, France
[4] URA CNRS 1974, Université Lyon 1, 69622, Villeurbanne cedex, France
[5] Laboratoire de Bio-Pédologie, ORSTOM - BP 1386 - Dakar - Sénégal
[6] Centre de coopération International en Recherche Agronomique pour le Développement
[7] Institut Sénégalais de Recherches Agricoles - Laboratoire National d'Elevage et de Recherches Vétérinaires



$$x_{i.} = \sum_{j=1}^{p} x_{ij} \quad x_{.j} = \sum_{i=1}^{n} x_{ij} \quad x_{..} = \sum_{ij}^{np} x_{ij}$$

$$f_{ij} = x_{ij}/x_{i.} \quad \bar{f}_j = x_{.j}/x_{..}$$

$$t_{ij} = x_{ij}/r_i \quad \bar{t}_j = x_{.j}/R \quad R = \sum_{i=1}^{n} r_i$$

- COA analyse a table F_0, the data of which are given by $(f_{ij} - \bar{f}_j)/\bar{f}_j$
- DCOA analyse a table T_0 the data of which are given by $(t_{ij} - \bar{t}_j)/\bar{t}_j$

Figure 1 : COA and DCOA calculation processes

Columns identify infectious agents and rows each village at each period. The villages where the size of sampled animals population was not sufficient for a certain period were excluded from the analysis. Let x_{ij} be the sum of the sampled animals presenting a rising titre of antibodies for a given agent *j* and a given [village x period] *i*, r_i all of the sampled animals (constant for whole agents), and $t_{ij} = x_{ij}/r_i$, the indicator of current infection. Decentred COrespondence Analysis (DCOA) was been used (described in details in Doledec et al., 1995). In this eigenvector technique which is similar to COrrespondence Analysis (COA) (Greenacre, 1993), classical row weights (written $x_{i.}$ in figure 1) built from the data themselves are replaced with new imported row weights r_i. DCOA make a singular value decomposition of the symetric matrice $[D_p T_0' D_n T_0]$, where D_p and D_n are respectively the diagonal matrices of column weights $x_{.j}/R$ and row weights r_i/R when COA make a singular value decomposition of $[D_p F_0' D_n F_0]$, where D_p and D_n are respectively the diagonal matrices of column weights $x_{.j}/x_{..}$ and row weights $x_{i.}/x_{..}$. The analysis gives an epidemiologic meaning to the ratio : t_{ij} represent the proportion of recently infected animals for a given [village x period], and $\bar{t}_j$ the mean of proportions for a given agent (Cochran, 1977).

II. RESULTS

All the analysis and graphical display have been computed with ADE4 (Chessel et al., 1995).

Agent	H1	H2	**H5**	**H6**	**H7**	H9	H11	MA	MD	**AD5**	**PI3**	**RSV**
p (χ^2)	0.292	0.07	**0.001**	**0.038**	**0.001**	0.391	0.336	0.085	0.897	**0.000**	**0.000**	**0.000**

Table I : multinomial tests of adjustment. p-values

Table I gives the p-values of χ^2 tests (25 df) between the observed (x_{ij}) and proportional to sampled population size (r_i) distributions. The 2 first eigenvalues represent 60 p. cent of the totality of variance of the table (sum of all the eigenvalues). Figure 2 plots the factorial scores of columns (2A) and rows (2B and 2C) projections on F1 and F2. To make the graphics easier to read, sampling units in each village (2B) and in each period (2C) have been linked together. In figure 2A , the zero point is the position of a theoritical agent for which the distribution of x_{ij} would be exactly proportional to r_i. For each variable, $\cos^2(\alpha_j) = d_j^2/\psi_j^2$ assess colinearity between projection and axis. $(x_{.j}/R)\psi_j^2/\lambda_k$ is the proportion of variance of row scores on axis *k* explained by *j*, with λ_k being the *k*th eigenvalue (table II). 6 variables strongly explain the 2 first axis. Each virus shows a relative independant position in comparison with the 2 others, when Pasteurella serotypes are quite correlated and form a redundant group. Spatial and temporel effects are obvious and interpretable. The first 2 periods (HS1 - RS) are opposed to the 2 others (CS - HS2) on F1, such as the villages of the 2 different areas. F2 separates 3 villages from the others, LP and SE on the one hand, and DI on the other hand, where infection rates for RSV and Pasteurella are higher respectively (Tillard et al., 1997).

Agent		H1	H2	**H5**	**H6**	**H7**	H9	H11	MA	MD	**AD5**	**PI3**	**RSV**
Relative contributions	F1	0.15	0.07	**0.04**	**0.00**	**0.38**	0.26	0.02	0.00	0.00	**0.36**	**0.85**	**0.17**
	F2	0.30	0.31	**0.77**	**0.63**	**0.44**	0.07	0.07	0.16	0.10	**0.00**	**0.04**	**0.35**
Absolute contributions	F1	0.02	0.01	**0.01**	**0.00**	**0.10**	0.04	0.01	0.00	0.00	**0.10**	**0.65**	**0.06**
	F2	0.06	0.07	**0.27**	**0.16**	**0.16**	0.01	0.01	0.04	0.01	**0.00**	**0.04**	**0.17**

Table II : absolute ($(x_{.j}/R)\psi_j^2/\lambda_k$) and relative ($\cos^2(\alpha_j) = d_j^2/\psi_j^2$) contributions of variables to axis.

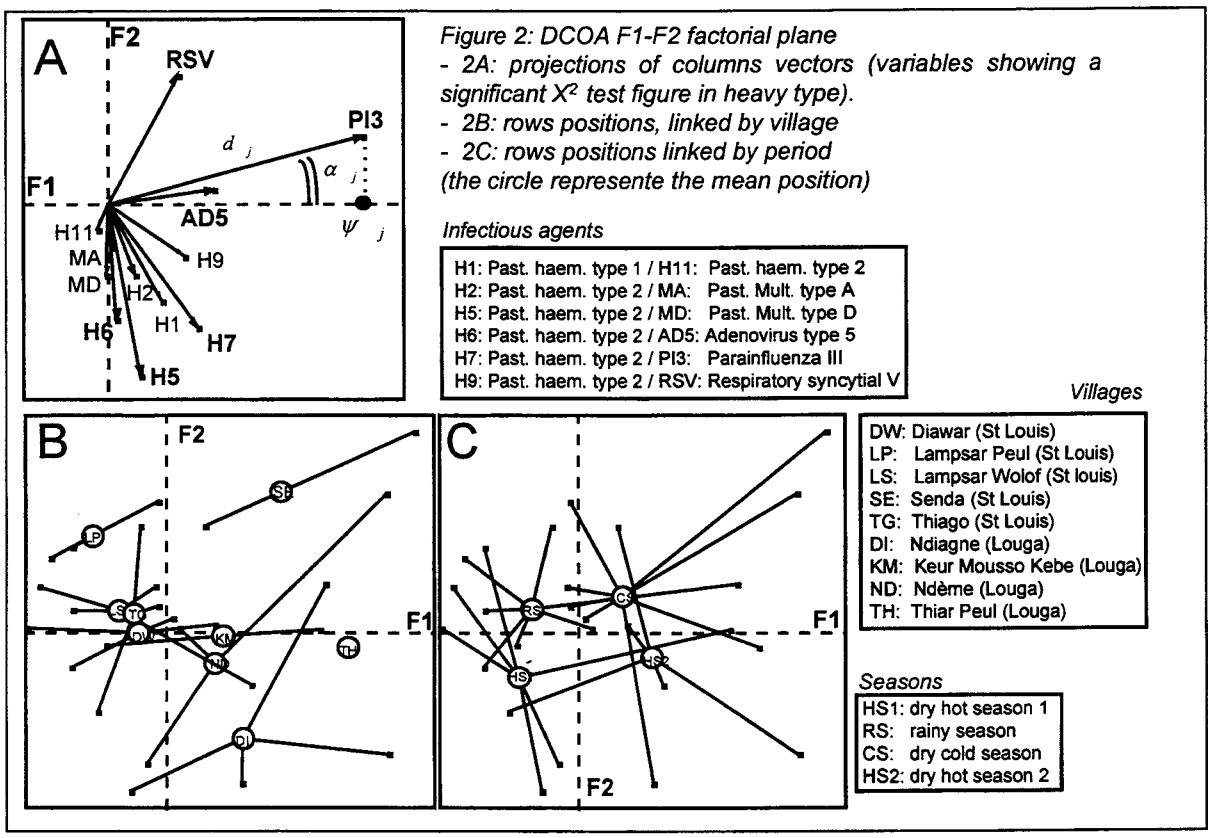


Figure 2: DCOA F1-F2 factorial plane
- 2A: projections of columns vectors (variables showing a significant X^2 test figure in heavy type).
- 2B: rows positions, linked by village
- 2C: rows positions linked by period
(the circle represente the mean position)

III. DISCUSSION ON METHODOLOGY

Faced with such a table of proportions, two others methods, Principal Components Analysis (PCA) of t_{ij}, and COA of x_{ij} (with classical row weights), could be considered first. PCA is not efficient from a mathematical point of view. A mean of proportions with different denominators is given by $\bar{t} = \sum_{i=1}^{n} x_{ij} \Big/ \sum_{i=1}^{n} r_i$ and is not an arithmetic mean (Cochran, 1977). For that reason, column centring usually made by PCA is not valid. COA of x_{ij} does not aim at the same purpose. Animals frequently present a rising titre for several agents within a same period. In such conditions, f_{ij} gives the proportion of animals showing a rising titre for a specific agent within whole of cases of rising titre for any of the agents. Animals presenting a decreasing or stable titre would be then ignored. In conclusion, DCOA is well adapted to the initial study aims. This technique allows the sampled population to be selected as the reference point (average point). It also permits the analysis of proportions with different denominators between rows. In a further stage of the analysis, a Canonical Correspondence Analysis (Ter Braak, 1986) will allow the matching of the proportions table with dependent variables describing the spatial and temporel organisation of the data and to compare villages and periods, even if the sampled population size is quite different.

Références

Chessel D., Doledec S., Olivier J.M., Thioulouse J. , 1995, ADE-4 Programme library, Multivariate analysis and graphical display of environmental data, User manual. 9 volumes. URA CNRS 1451, University of Lyon 1, http://biomserv.univ-lyon1.fr/ADE-4.html.
Cochran W.G., 1977, Sampling techniques, Wiley J. & sons (Ed.), New York, 428 p.
Desoutter D., 1994, Les pneumopathies des petits ruminants au Sénégal: étude séro-épidémiologique et bio-géographique, Thèse de Doctorat d'Université, Université Paris XII Val de marne - Créteil, 189 p. et annexes.
Doledec S., Chessel D., Olivier J.M., 1995, L'analyse des correspondances décentrée: application aux peuplements ichtyologiques du Haut Rhône, Bull. Fr. Pêche Piscic., 336, 29-40.
Greenacre M., 1984, Theory and applications of Correspondence Analysis, Academic Press, London, 364 p..
Ter Braak C.J.F., 1986, Canonical Correspondence Analysis : a new eigenvector technique for multivariate direct gradient analysis, ecology, 69, 69-77.
Tillard E., Faye B., 1997, Analyse d'un tableau de taux : Exploration multidimensionnelle de la circulation des agents infectieux pneumotropes chez les petits ruminants au Sénégal, *ftp ://biom3.univ-lyon1.fr/pub/mac/ADE/ADE4/DocThemPDF/ADE-Sec-8.3.pdf.*