Epidémiol. santé anim., 1997, 31-32

THE USE OF THE CANONICAL CORRESPONDENCE ANALYSIS IN EPIDEMIOLOGY

Faye B.[1], Lescourret F.[2], Tillard E.[2]

Les relations entre les fréquences de 6 pathologies mammaires (traumatisme, troubles physiques, fonctionnels, congestion mammaire, mammites cliniques et subcliniques) et l'ensemble des pratiques et des conditions d'élevage (hygiène, pratiques de traite, bâtiment d'élevage, données individuelles agrégées) dans 187 élevages-années sont étudiées en s'appuyant sur une analyse canonique des correspondances (ACC). Cette méthode relève de la technique d'analyse des correspondances dans laquelle les axes représentant les profils sanitaires sont l'expression des combinaisons linéaires des variables décrivant les pratiques de management. Cette méthode est la généralisation d'une régression linéaire de deux jeux de variables, l'un étant explicatif du second. Pour cela, on procède de la manière suivante: (i) le premier jeu de données (P) comprend les fréquences des 6 pathologies mammaires dans 187 élevages bovins laitiers, (ii) le second (X) comprend les modalités des variables décrivant les pratiques de management dans ces mêmes exploitations laitières.
L'ACC procède en deux étapes: dans un premier temps, le tableau P est projeté dans l'espace vectoriel (combinaisons linéaires) de X, puis, dans un second temps, une analyse factorielle des correspondances de ces projections (AFCp) est réalisée. Le diagramme produit par l'ACC permet de visualiser la distribution des pathologies mammaires parmi les variables de management les plus explicatives. Le rapport des variances expliquées par l'AFC simple de P et par l'ACC de P sur X reflète le pourcentage de variance expliquée des pathologies mammaires par les pratiques de management.

INTRODUCTION

In many cases, observational studies aim to assess the relationships between two groups of variables: the first consists of the multiple dependant variables of direct interest (such as disease occurrences); the second is composed of variables that are supposed to influence the variables of the first group. Their relationships are fundamentally asymetric. The simultaneous analysis of such group of variables -taking in account this asymetric relationship- has been proposed by using Canonical Correspondence Analysis (CCA). It's a multivariate analysis based on a generalization of Rao's (1964) principal component analysis with instrumental variables (e.g. explained variables).
To our knowledge, CCA has been used mainly in ecological studies on species-environment relationships (Lebreton *et al,* 1988; Birks and Austin, 1992) based on the simultaneous analysis of two datasets: the first includes occurrence of different species, the second describes environmental conditions. The use of this method in epidemiological studies is more recent (Lescourret and Faye, 1991).

MATERIAL AND METHODS

Data were collected during a 4-yr study (1986-1990). The purpose, strategy and methodology have been described elsewhere (Faye *et al,* 1989). A total of 4129 dairy cows (holstein breed) having 8945 lactations from 47 intensive dairy farms in Brittany (France) were followed. The herd-year was the observation unit (n=187). All the informations concerning health, milk production, milk bacteriology, farming practices and quality data were organized in a relational database under ORACLE (Lescourret *et al,* 1993).
Initially udder-health status for each herd-year was measured with 6 variables: (1) incidence risk of traumatic udder disorders (UDTR) including teat crushing, teat cuts and udder injuries, (2) incidence risk of physical udder disorders (UDPHY) including chaps, warts, cracks, vaccinias and mammary tumors, (3) incidence risk of functional udder disorders (UDFUN) including blood in the milk, milk retention and agalactia, (4) incidence risk of congestive disorders (UDCON) such as udder edema after calving and udder congestion, (5) incidence risk of clinical mastitis, and (6) prevalence risk of subclinical mastitis (i.e., SCC greater than 400 x 10^3 ml $^{-1}$ (SUBM).
This first dataset (P) of 187 herd-years (rows) by 6 variables describing udder-health status (columns) was studied with a simple factorial correspondence analysis (SFCA) to compare the results with canonical correspondence analysis (see later).
After step-wise analysis,13 qualitative variables among 47 were considered: type of quarantine (QUARANT: 5 levels), disinfection of animal housing (DISAH, 2 levels), wet patches in the litter (WETLIT, 2 levels), type of teat dipping (DIP, 5 levels), number of udder-towel per herd (UDDTROW, 5 levels), mean milking time per cow (MILKTIM, 3 levels), claw rinsing (RINS, 2 levels), housing orientation (ORIENT, 5 levels), shelter at pasture (SHELPAST, 2 levels), type of housing (HOUSTYP, 3 levels), rain protection (RAIN, 3 levels), mean daily milk production (MILKPROD, 3 levels), mean loss of fat score after calving (LOSSFAT, 3 levels).
The relationships between the 6 outcomes and the 13 covariates were studied using a canonical correspondence analysis (CCA) (Chessel *et al,* 1987; Ter Braak, 1986). This method can be considered as a

[1] CIRAD-EMVT Campus international de Baillarguet. BP 5035. 34032 Montpellier cedex, France
[2] Laboratoire d'écopathologie. INRA-Theix, 63122 St Genes Champanelle, France

generalization of regression to two sets of variables, the first being explained by the second. This method proceeds as follows:(1) the outcome P includes 6 columns for the risks of the 6 udder-health variables and 187 herd-year rows, (2) the explanatory variables X includes the levels oberved for the covariates describing farming management.

Formally, CCA proceeds in two steps: (1) the dataset P (udder health status) is projected in the linear combinations of covariates defined by the dataset X (management practices), and (2) a factorial correspondence analysis of these projections (called here PFCA to distinguish it from the previous SFCA) is performed. So, the previous SFCA decomposed the total variance of P (udder health status) whereas the CCA decomposed only the projected variance of P on X (management practices). In other words, we determined the non-correlated linear combinations of the explained variables which maximized the variance of the dependant variables group (Obadia, 1978).

If management practices totally explain udder-health status, then the CCA and the previous SFCA (see below) would give the same result. The ratio of variances (i.e. sum of the eigen values) computed by the SFCA on table P and by the CCA on tables P and X respectively, express the global explanation of udder-health status by management practices (Yoccoz, 1988). This ratio is called the "ratio of canonical correlation" (or "redundancy index" by Israels, 1984).

The interpretation of the CCA factors is based on their correlations with the levels of the covariates (Lebreton *et al,* 1988). The overall coorelation (h^2) between an expalanatory variable and a factor is the sum of the correlations between each level of the variable and the factor (r_i^2), weighted by level of occurrence Qi ($h^2 = S(1-Q_i)r_i^2$). This is interpreted as the square of correlation coefficient and is assessed for significance (Saporta, 1990). If an explanatory variable is considered significantly associated with CCA factors then the individual r_i^2 are assessed both in terms of direction (positive and negative correlations) and size of relationship.

RESULTS

The yearly per herd incidence risk (mean and sd) of non-infectious udder disorders was low: 0.64 ± 1.23 for UDTR, 0.47 ± 1.25 for UDPHY, 0.94 ± 1.65 for UDFUN and 0.96 ± 2.1 for UDCON. The mastitis incidence risk was 14.1 ± 8.6 and the prevalence risk of high SCC was 24.0 ± 11.5. The maximum value across the herd-years was respectively 7.1 (UDTR), 11.4 (UDCON), 6.5 (UDFUN), 11.5 (UDPHY), 53.5 (MAS) and 63.3 (SUBM). The minimum value was nil in all cases.

The six outcomes of the udder-health status were represented by 5 SFCA factors (table I). The only significant variable in the first factor (F1) was congestive udder disorders (UDCON). F1 explained 28% of the total variation) and UDCON accounted for 85.7% of the variance of F1. F2 consisted of 3 variables, clinical mastitis (MAS) and physical disorders (UDPHY) having positive coefficients and high SCC (SUBM) having a negative coefficient. F2 explained 23% of the total variation, with all 3 significant variables each explaining approximately 30% of the variance of F2. On this factor, MAS and SUBM were in opposition to each other according to their factorial coordinates (coefficients) on F2.

Table I

Analysis of the dataset P (occurrence of udder disorders in 187 herd-years): relative contribution (in %) and *factorial coordinates* (coefficients) of outcomes of udder-health status significantly correlated (p <0.01) to the factors of the Simple Factorial Correspondance Analysis (SFCA)

	Relative contribution to factors from SFCA *(Proportion in % of the explained variance)* [b]				
udder disorders	F1 *(28)*	F2 *(23)*	F3 *(19)*	F4 *(16)*	F5 *(14)*
trauma	*a*	*a*	*a*	59.4 *(+1.58)*	24.8 *(+0.92)*
physical disorders	*a*	29.6 *(+1.57)*	50.1 *(+1.81)*	*a*	*a*
functional disorders	*a*	*a*	27.8 *(-0.95)*	*a*	57.6 *(-1.16)*
congestion	85.7 *(+2.05)*	*a*	*a*	*a*	*a*
clinical mastitis	*a*	31.4 *(+0.29)*	*a*	23.1 *(-0.21)*	*a*
subclinical. mastitis	*a*	28.6 *(-0.21)*	*a*	*a*	*a*

a. Contribution to the factor was non-significant (P <0.01)

b. The sum of the explained variance by the factors F1 to F5 is 100%

The CCA of P,X identified two combinations of covariates correlated to F1 and to F2. The combination of no quarantine and loss of fat after calving was unfavorable for congestive udder disorders while the combination no introduction + low or average loss fat was unfavourable for mastitis and physical disorders (table II). For all types of non-infectious udder disorders (except physical) the most unfavourable combination of covariates was individual udder towel + cubicles + low milk production. Inversely, for clinical and subclinical mastitis, the combination no udder towel + loose housing + high milk production was quite unfavourable. Overall this model explained 12% of the total variance.

Table II
Most-contributive combinations of covariates to factors of the CCA of P, X and mean yearly incidence risk of udder disorders (in %) for each combination

combination of covariates	udder disorders incidence				
First factor of P, X_j	mas		udphy		udcon
quarant1+lossfat3	12.2[a]		0.15[a]		1.49[a]
quarant5+ lossfat2 or lossfat1	14.5[b]		0.80[b]		0.64[a]
Second factor of P, X_j	mas	subm	udtr	udfun	udcon
uddrtow1+houstyp2+milkprod3	17.3[a]	30.7[a]	0[a]	0.85[a]	0[a]
uddrtow5+houstyp1 or 3+milkprod1	13.7[b]	12.9[b]	0.94[b]	1.5[b]	0.83[b]

Test value (Kruskall-Wallis test) per column within factor: [a-b]: $P < 0.05$

DISCUSSION

In the frame of this paper, only statistical aspects will be discussed. Biological aspects has been discussed elsewhere (Faye *et al.,* 1997). By using an average measurement of relationships between the variables of P (diseases) and X (farming practices), Canonical Correspondence Analysis emphasized the natural dissymetry of these relationships. Thus, CCA extends the MANOVA concept to datatables playing a asymetric role (Takeuchi *et al,* 1982; Sabatier *et al,* 1989). The aim of the CCA method is maximization of the explained variance of the dependant variables by a linear combination of dependant variables , having the highest multiple correlation with covariates (Obadia, 1978). According to Lebreton *et al.,* (1991), this partly bridges the gap between descriptive factorial method ("l'analyse de données" of french statisticians) and the standard approaches of multivariate analysis. The development of more sophisticated inference procedures in more complex cases are expected. They will considerably improve the level of flexibility so far achieved.

BIBLIOGRAPHY

Birks H.J.B.,Austin H.A., 1992. An annoted bibliography of canonical correspondence analysis and related constrained ordination methods (1986-1991). Botanical Institute, All-Gaten 41, N-5007 Bergen, Norway, 1-29

Chessel D., Lebreton D., Yoccoz N., 1987. Propriétés de l'analyse canonique des correspondances: un exemple en hydrobiologie. Rev. Stat. Appl., 35, 55-72

Faye B., Barnouin J., Lescourret F. , 1989. Objectifs principaux et stratégie de l'enquête écopathologique Bretagne sur la vache laitière. Epidémiol. Santé Anim., 15, 23-31

Faye B., Lescourret F., Dorr N., Tillard E., McDermott B., McDermott J., 1997. Interrelationships between herd management practices and udder health status using canonical correspondence analysis. Prev. Vet. Med. (Accepted)

Israels A.Z., 1984. Redundancy analysis for quantitative variables. Psychometrika, 49, 331-346

Lebreton J.D., Chessel D., Richardot-Coulet M., Yoccoz N., 1988. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. Acta Oecol., 9, 137-151

Lebreton J.D., Sabatier R., Banco G, Bacou A.M., 1991. Principal components and correspondence analyses with respect to instrumental variables: an overview of their role in studies of structure activity and species-environment relationships. In: J. Devillers and W. Karcher (Ed). Applied multivariate analysis in SAR and environmental studies, EEC, Brussels, Belgium, 85-114

Lescourret F., Faye B., 1991. Stratégie statistique du laboratoire d'écopathologie. Epidemiol. Santé Anim., 2, 103-115

Rao C.R., 1964. The use of the interpretation of principal component analysis in applied research. Sankhya, Ser. A, 26, 329-359

Obadia J., 1978. L'analyse en composantes explicatives. Rev. Stat. Appl., 26, 5-28

Sabatier R., Lebreton J.D., Chessel D., 1989. Principal component analysis with instrumental variables as a tool for modelling composition data. In : R.Coppi and S. Balasco (Ed), Multiway data analysis. Elsevier, Holland, Amsterdam, 341-352

Saporta G., 1990. Probabilités, analyse des données et statistiques. Editions Technip, Paris, 493 pp.

Takeuchi K., Yanai H., Mukherjee B.N., 1982. The foundations of multivariate analysis. A unified approach by means of projection onto linear subspaces. J. Wiley and sons, New-York, USA, 458 pp.

Ter Braak C.I.J., 1986. Canonival correspondence analysis: a new eigenvector technic for multivariate direct gradient analysis. Ecology, 67, 1167-1179

Yoccoz N.G., 1988. Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse, Université de Lyon, France.