Epidémiol. Santé anim., 1984, n° 6, 32-60

UTILISATION DES METHODES D'ANALYSE DES DONNEES POUR L'ETUDE DES MALADIES D'ELEVAGE. APPLICATION AU PORC.

F. MADEC et J. JOSSE**

RESUME : Les maladies à étiologie complexe des élevages porcins ont une incidence économique de grande importance. Leur reproduction en station expérimentale est très difficile ; les médications sont coûteuses et donnent des résultats aléatoires. Le dépistage des conditions associées à leur éclosion dans les élevages suppose la mise en place d'un protocole d'observation détaillée sur une période assez longue. Le nombre de variables enregistrées est alors considérable et le traitement de l'information requiert des moyens adaptés. Les méthodes de l'analyse des données constituent une solution de choix pour le dépouillement de ces enquêtes. Cependant, il est nécessaire d'assortir ces méthodes des logiciels d'accompagnement en vue, d'une part, de la préparation des données à soumettre à l'analyse et, d'autre part, de l'interprétation des résultats obtenus. Par ailleurs, les études sur les maladies d'élevage supposent l'adoption d'une démarche générale d'observation, d'interprétation et d'action pour la mise en oeuvre des mesures préventives (correction des facteurs de risque).

SUMMARY: The diseases of multifactorial etiology cause great economic losses in pig production. Experimental trials are difficult to carry on and they are hardly conclusives, while chemiotherapy is expensive and gives irregular results. In order to point out the environmental conditions regularly associated with the troubles, a detailed review of the herd management is required. So a great number of variables are involved and the statistical study becomes complex. The multidimensional data analyse methods are of good opportunity. Although, it is necessary to get an oppropriate complementary software to prepare the data board and to help rendering the results obtained. Further, ecopathological studies lead to follow a general proceeding of attentive observation, analysis and action in order to improve the failing parameters among the risk factors.

*

* *

* *Ministère de l'Agriculture, Direction de la Qualité, Services vétérinaires, Station de Pathologie Porcine, B.P. n° 9, 22440 Ploufragan.*

PREMIERE PARTIE :

SITUATION DU PROBLEME ET METHODES D'ANALYSE

- Introduction -

L'observation quotidienne des élevages intensifs de porcs permet d'établir une classification schématique des phénomènes pathologiques qui s'y développent. Sur la base des données expérimentales obtenues au laboratoire et des données cliniques et épidémiologiques observées sur le terrain, on peut ranger actuellement la plupart des maladies du porc en deux grandes catégories :

- les maladies qui font intervenir de manière exclusive un agent étiologique donné. Ces affections présentent en commun un caractère contagieux rendant possible leur propagation à l'intérieur d'un élevage et d'un élevage à l'autre (maladies épizootiques). Des maladies virales entrent essentiellement dans cette catégorie : gastro-entérites contagieuses, grippe porcine, maladie d'Aujeszky, fièvre aphteuse et pestes porcines;

- des maladies trouvant dans certains troupeaux des conditions favorables à leur éclosion puis à leur persistance et prenant l'allure de maladies enzootiques : on les dénomme couramment "maladies d'élevage". Elles mettent ou non en jeu des agents microbiens mais, à la différence des précédentes, ces agents n'exercent ici leur effet pathogène que dans des conditions de milieu particulièrement propices. La reproduction expérimentale de l'intégralité des signes cliniques et des lésions observées spontanément dans les élevages est difficile. Leur déterminisme est complexe et elles entrainent, en dépit de manifestations cliniques généralement peu spectaculaires, les pertes économiques les plus lourdes pour l'élevage porcin car elles pénalisent les résultats techniques (croissance, efficacité alimentaire - TILLON et al., 1980). Les exemples sont très nombreux : troubles digestifs du sevrage, troubles respiratoires, troubles de la mise-bas, troubles urinaires, troubles de la reproduction...

Les moyens de contrôle et de prévention des maladies du porc vont considérablement diverger selon le groupe d'appartenance. Ainsi, on mettra en oeuvre une lutte directe contre l'agent causal dans les maladies épizootiques (sérum, vaccin, protection sanitaire), alors que pour les maladies d'élevage la démarche est à la fois plus complexe et plus lente. L'importance actuelle de ces maladies dans les élevages atteste d'un certain manque d'information des professionnels. Une stratégie cohérente est requise et elle ne peut reposer que sur la connaissance approfondie du milieu.

La présente publication rend compte de l'intérêt et des contraintes du recours aux méthodes scientifiques de l'analyse des données pour l'appréhension de ces conditions de milieu qui favorisent les maladies d'élevage.

I. SITUATION DU PROBLEME

L'étude des maladies d'élevage, doit être conduite dans les élevages eux-mêmes ce qui représente une entreprise difficile. En effet, dès lors que l'on désire non plus simplement décrire un phénomène global, par exemple des résultats de productivité faible mais en trouver l'origine, l'explication ou les circonstances d'apparition, des difficultés surgissent de toutes parts :

- les contours du syndrome clinique sont en général mal délimités, les manifestations pathologiques sont très fluctuantes selon les élevages ; de plus, elles évoluent dans le temps au sein d'un même troupeau ;

- la plus grande partie des travaux ayant été réalisés à propos des maladies contagieuses, les rapports bibliographiques relatifs à certains syndromes sont peu nombreux et concernent le plus souvent des études fragmentaires ;

- l'épreuve expérimentale préalable ou contemporaine à l'étude épidémiologique n'est pas toujours réalisée ou bien sa signification est douteuse. Il est en effet très difficile de reproduire en conditions contrôlées, la situation d'un élevage de production. Et même en supposant que cela soit possible, il ne s'agirait que d'une seule unité observation, sans groupe témoin ;

- le déterminisme complexe des maladies d'élevage oblige à élargir l'éventail du recueil des données afin de s'assurer d'une bonne exhaustivité. En conséquence, on aboutit très rapidement à un alourdissement de la collecte des informations, de leur saisie et de leur traitement. En outre, les données recueillies doivent être fiables et objectives ce qui suppose l'élaboration d'un protocole approprié ;

- l'approche statistique des données est essentiellement analytique en raison de la spontanéité d'apparition des phénomènes étudiés. En outre, elle est comparative d'où la nécessité de disposer d'un nombre suffisant d'observations pour effectuer les comparaisons.

Toutes ces considérations mettent en relief l'importance des moyens qu'implique une telle opération. Il ne peut s'agir que d'une équipe multidisciplinaire dont les membres (statisticiens-informaticiens, spécialistes de la santé...) travaillent en collaboration tout au long du déroulement de l'étude. Toutes ces raisons ne sont peut être pas étrangères au faible nombre d'études du même type qui ont pu être conduites à leur terme à ce jour.

II. CHOIX DE LA METHODE DE TRAITEMENT DE L'INFORMATION

Le traitement de l'information et le choix des méthodes sont une étape importante de l'étude :

- la difficulté majeure rencontrée dans l'analyse des données en épidémiologie est d'abord le très grand nombre de variables disponibles (plusieurs centaines) ;

- les données sont enregistrées "in situ" dans les élevages. Les phénomènes pathologiques étudiés surviennent spontanément sans aucun artifice de la part de l'expérimentateur et indépendamment de la volonté du maître d'oeuvre. En cela, ces études sont à distinguer de l'expérimentation basée sur un plan expérimental obéissant aux schémas statistiques habituels et un protocole d'intervention (moment de l'inoculation, dose d'inoculum...) volontairement conçu par l'expérimentateur et qui ne se rencontre pas naturellement dans les élevages ;

- en raison de l'importance accordée à la qualité des données, ces dernières sont rangées selon un certain ordre mais celui-ci n'engage nullement à l'utilisation d'une méthode statistique donnée. La procédure de traitement reste libre en aval car il n'existe pas d'option précise a priori sur la stratégie à suivre. Réciproquement, la très grande rigidité de certaines méthodes statistiques va inévitablement les exclure ;

- le but de l'enquête écopathologique est de faire apparaître, au terme d'une étude comparative méthodique et approfondie, les circonstances majeures qui sont régulièrement associées à l'existence d'un phénomène pathologique ;

- ces circonstances ou variables explicatives doivent, en toute logique, entretenir avec le problème considéré des relations privilégiées. Les méthodes mathématiques qui permettent en toute rigueur de mettre en évidence les liaisons qui existent dans un tableau complexe seront adaptées à l'objectif. On doit signaler ici que la nature biologique et fluctuante du "substrat" et surtout l'absence de contrôle* de plusieurs variables, rend utopique et illusoire la recherche d'une signification statistique ou d'une quelconque précision* comme c'est le cas dans la plupart des essais. Les objectifs et l'attitude du maître d'oeuvre sont pragmatiques : la méthode vise la détermination d'indicateurs de risque dans le but d'intevenir sur ces derniers.

Les méthodes de l'analyse des données sont à même de satisfaire les exigences venant d'être énumérées. Encore faut-il disposer des moyens de calculs appropriés.

III. LES METHODES DE L'ANALYSE DES DONNEES

A. Présentation des méthodes

Parmi les nombreuses méthodes disponibles (LEBART et al., 1979), trois sont particulièrement adaptées au traitement d'une enquête :

- l'analyse en composantes principales (A. C. P.) LEBART, FENELON (1975) ;

- l'analyse factorielle des correspondances (A. F. C.) BENZECRI (1979); ESCOFIER (1979) ;

---

* au sens de la statistique.

- les classifications automatiques, parmi lesquelles la classification ascendante hiérarchique (C. A. H.) JAMBU (1976).

Toutes ces méthodes ont un objectif commun : partant de n sujets sur lesquels ont été mesurées k variables, elles recherchent la visualisation des relations privilégiées existant dans le tableau des données. Elles sont par essence analytiques et leur vocation première est descriptive. Mais lorsque l'on dispose de ce moyen et qu'en plus on imagine une stratégie appropriée pour son utilisation, on peut valablement quitter le stade de la simple description pour déboucher sur la recherche de relations circonstancielles ou même de causalité.

Pour chacune des études écopathologiques menées à ce jour par notre équipe, l'A. F. C. et la classification hiérarchique ascendante ont été utilisées, en complément l'une de l'autre.

En raison du caractère qualitatif d'une partie importante des informations recueillies, la méthode de base reste l'A. F. C. (BENZECRI, 1964 ; FENELON, 1981). La possibilité d'une représentation simultanée des variables et des observations sur le même graphique de l'analyse (FENELON, 1981) facilite la compréhension de ce dernier par rapport aux résultats de l'A. C. P. La réalisation de ces analyses nécessite une bonne chaîne de programmes de traitement qui comportera un logiciel pour la mise en forme des données, un programme complet de calcul de l'analyse factorielle avec contributions absolues et relatives, et également des programmes d'aides à la visualisation et à l'interprétation. Par ailleurs, un programme de classification hiérarchique sur les facteurs de l'analyse est très utile en tant que complément de l'A. F. C. L'interprétation d'une analyse est délicate et requiert une bonne connaissance du principe de la méthode.

B. L'analyse factorielle des correspondances (A. F. C.)

1. Notions générales et définitions

Nous nous contenterons de donner une représentation sommaire, sans reprendre tous les développements mathématiques liés à l'analyse des correspondances*.

Le tableau de données est un croisement de nombres positifs dont les colonnes représentent un caractère à plusieurs modalités exclusives et les lignes un autre caractère. A l'intersection de la ligne i et de la colonne j se trouve l'élément noté K (i, j) qui correspond au nombre de fois que la modalité i et la modalité j se trouvent simultanément prises par les individus de la population totale étudiée.

On a l'habitude de noter :

. le tableau $K_{ij} = \{K(i,j) \; ; \; i \in I, \; j \in J\}$

avec I = ensemble des modalités du 1er caractère
avec J = ensemble des modalités du 2ème caractère.

---

* Le lecteur qui craindrait d'être rebuté par les formules mathématiques des pages suivantes pourrait se reporter à l'article de J. JOSSE, p. 26 à 28 et continuer la lecture à partir de la page 43.

. $K(i) = \sum_{j \in J} K(i,j)$; $K(j) = \sum_{i \in I} K(i,j)$

. $K = \sum_{\substack{i \in I \\ j \in J}} K(i,j)$ masse totale

. $f_{IJ} = \{f_{ij} = K(i,j)/K \; ; \; i \in I, \; j \in J\} \subset R_{I \times J}$

. $f_I = \{f_i = K(i)/K \; ; \; i \in I\} \subset R_I$

. $f_J = \{f_j = K(j)/K \; ; \; j \in J\} \subset R_J$

Les tableaux f_{IJ}, f_I, f_J, calculés à partir du tableau K_{IJ}, vérifient les axiomes des lois de probabilités ; ainsi

. f_{IJ} est la loi produit d'ensembles finis I et J ; f_I et f_J sont les lois marginales de la loi f_{IJ}. A partir de f_{IJ} on définit les lois conditionnelles suivantes :

$$f_I^j = \{ f_i^j = f_{ij}/f_j \; ; \; f_j \neq 0 \; ; \; i \in I \}$$

$$f_J^i = \{ f_j^i = f_{ij}/f_i \; ; \; f_i \neq 0 \; ; \; j \in J \}$$

. f_J^i est le profil de l'élément $i \in I$ sur J

En fait, le tableau analysé ne sera pas le tableau brut des données initiales $(K_{IJ} = K(i,j) \; i \in I, \; j \in J)$ mais le tableau des profils que l'on vient de définir. L'élément i sera caractérisé non pas par la ligne brute $\{K(i,j) \; ; \; j \in J\}$ de total $K(i)$ mais par son profil $\{f_J^i, \; j \in J\}$ chaque terme étant déduit de la ligne brute :

$$f_j^i = K(i,j)/K(i)$$

le total de la ligne étant ici I.

Ainsi, deux éléments ayant les mêmes coordonnées à un facteur multiplicatif près, auront le même profil.

On considère le nuage de points $N_J(I)$ défini de la manière suivante :

$$N_J(I) = \{f(f_J^i, f_i) \; i \in I\} \subset R_J$$

et le nuage de points N_I (J) défini de la même façon par :

$$N_I\ (J) = f\ (\overset{j}{I},\ f_j)\ j \in J\} \subset R_I$$

avec l'effectif de I ou cardinal (noté card I) = n

et l'effectif de J ou cardinal (noté card J) = m.

S'il était possible d'observer ce nuage de points dans l'espace à m dimensions, la structure de l'ensemble des observations serait immédiatement accessible : les observations aux profils de descriptions proches se situant dans la même zone, il appartient alors au spécialiste de découvrir la raison de leurs ressemblances.

En raison de l'impossibilité de visualiser les phénomènes dans les espaces supérieures à trois, il est nécessaire d'utiliser un artifice pour étudier ces structures. On détermine alors un nouveau système d'axes de références et on n'étudie les projections du nuage que sur les nouveaux axes les plus pertinents.

2. Métrique du X^2 et distances distributionnelles

On définit une distance euclidienne dite du X^2 de la façon suivante : soit l'espace des profils des observations sur J, la distance distributionnelle entre i et i' de I s'écrit ainsi :

$$||f_J^i - f_J^{i'}||^2_{f_J} = \sum_{j \in J} (f_j^i - f_j^{i'})^2/f_j$$

f_J = centre du N_J (I)

de même, la distance distributionnelle entre j et j' de J est

$$||f_I^j - f_I^{j'}||^2_{f_I} = \sum_{i \in I} (f_i^j - f_i^{j'})^2/f_i$$

la distance d'un élément i au centre du nuage N_J (I) est donc :

$$\rho^2\ (i) = ||f_J^i - f_J||^2_{f_J} = \sum_{j \in J} (f_j^i - f_j)^2/f_j$$

En analyse des correspondances, le choix de la métrique du X^2 de centre f_j est lié au principe d'équivalence distributionnelle c'est-à-dire : pour deux éléments j et j' tels que leurs profils soient identiques, si on substitue aux colonnes j et j' une colonne j_s telle que $f_{ij} = f_{ij} + f_{ij'}$, $\forall i \in I$, alors la distance entre les éléments de I n'est pas modifiée.

Selon les données analysées, le nuage N_J (I) sera plus ou moins dispersé autour de son centre de gravité f_j. Cette dispersion peut être mesurée par l'inertie du nuage M qui est la somme, pour tous les éléments du nuage, du produit du carré de la distance au centre du nuage par la masse de l'élément.

D'où

$$M^2\{ N_J\ (I)\} = \sum_{i \in I} f_i \rho^2\ (i)$$

et

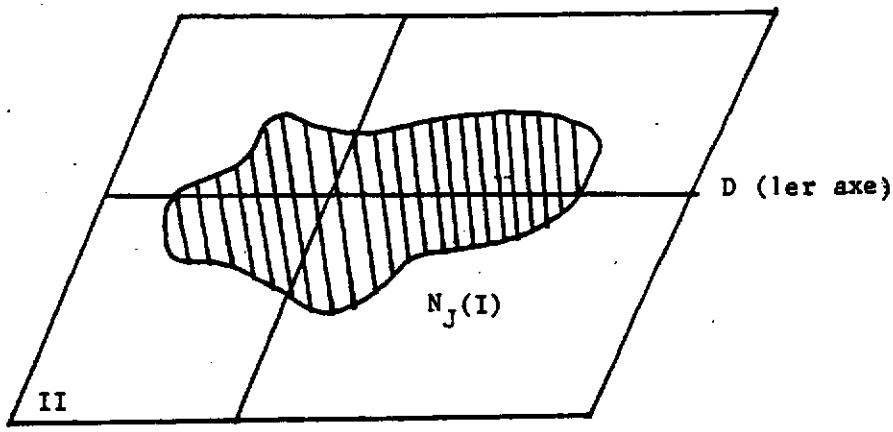
$$M^2\{ N_I\ (J)\} = \sum_{j \in J} f_j \rho^2\ (j)$$

on montre que $M^2\{ N_J\ (I)\} = M^2\ \{ N_I\ (J)\}$

3. Axes factoriels et facteurs

On cherche une représentation du nuage N_J (I) qui, en dimension aussi faible que possible, soit aussi fidèle que possible. Pour cela on projette le nuage N_J (I) sur un sous-ensemble linéaire L (droite ou plan) passant par son centre de gravité f_J. C'est la projection du nuage sur L qui constituera la représentation approchée désirée. Pour fixer la pensée, considérons le nuage N_J (I) ci-dessous, aplati au voisinage d'un plan II et plus allongé dans la direction d'une droite D de ce plan. Il est naturel de rapporter l'espace à de nouveaux axes.

Le premier axe sera porté par D, le deuxième par la perpendiculaire à D en f_J dans le plan II. Les coordonnées des profils f_j^I sur ces axes sont les facteurs.



Le problème est donc de chercher le sous-espace L passant par le centre de gravité f_J et rendant maximale l'inertie de N (I) le long de L.

Le programme d'analyse des correspondances calcule les nouveaux repères ou axes factoriels ainsi que, pour les deux ensembles I et J, les facteurs (F_α, G_α) associés au nombre λ_α (valeur propre associée) réel positif qui décroît de facteur en facteur.

On a donc :

$$F_1 = (F_1\ (i),\ i \in I)\ ;\ G_1 = (G_1\ (j)\ ;\ j \in J)\ ;\ \lambda_1$$

$$F_\alpha = (F_\alpha\ (i),\ i \in I)\ ;\ G_\alpha = (G_\alpha\ (j)\ ;\ j \in J)\ ;\ \lambda_\alpha$$

$$F_N = (F_N\ (I),\ i \in I)\ ;\ G_N = (G_N\ (j)\ ;\ j \in J)\ ;\ \lambda_N$$

Où N = inf (Card I - 1, Card J - 1) et $\lambda N < \ldots < \lambda_\alpha < \ldots < \lambda 1$

F_α (i) = valeur du facteur de rang α au point $i \in I$

G_α (j) = valeur du facteur de rang α au point $j \in J$

Dans la pratique, les premiers axes factoriels sont les plus importants puisqu'ils sont édités dans l'ordre décroissant de leur variance et l'étude peut souvent se limiter aux quatre premiers. La détermination du nombre d'axes étant faite à partir d'un spectre des valeurs propres.

Enfin, pour bien interpréter les résultats, il faut disposer d'une indication sur la qualité de la représentation d'un point sur un facteur. Dans ce but on définit, d'une part, la contribution relative du facteur au point i qui se calcule de la façon suivante :

$$F^2_{\alpha}(i)/\rho^2(i)$$

et, d'autre part, la contribution absolue du point i au moment d'inertie λ_{α} qui permet de connaître les éléments participant le plus à la déformation du nuage selon l'axe factoriel α soit

$$f_i F^2_{\alpha}(i) \text{ et naturellement } \sum_{i \in I} f_i F^2(i) = \lambda_{\alpha}$$

C. Classification ascendante hiérarchique à partir d'un tableau de facteurs

1. Utilisation

L'analyse des correspondances a le mérite de faire figurer sur un même graphique deux ensemble I et J ; cependant, les praticiens de l'analyse des données doivent souvent fixer en un système rigide de classe la vision globale et spatiale que fournit l'analyse factorielle ; c'est pourquoi on a recours à la classification automatique (JAMBU et LEBEAUX, 1978 ; BRUYNOOGHE, 1978 ; BUARD, 1980).

On peut effectuer directement une classification ascendante hiérarchique sur l'un des ensembles I ou J à partir d'un tableau de correspondances sur I x J, mais il est beaucoup plus facile de le faire à partir des coordonnées factorielles fournies par une analyse des correspondances en sachant que la distance du X^2 entre deux éléments de J dans l'espace I et deux éléments de I dans l'espace J est égale à la distance euclidienne usuelle entre ces deux mêmes éléments dans l'espace des facteurs.

Une méthode de classification est tributaire de trois paramètres :

- le calcul d'une distance entre éléments,
- l'affectation d'une masse à chaque élément,
- le calcul d'une distance entre classes d'éléments qui dépend du critère d'agrégation.

Ici le tableau des données est constitué des facteurs associés aux éléments à classifier (de I par exemple) dont on a retenu un nombre A interprétable :

$$F_{\alpha}(i) \; ; \; i \in I, \alpha = \{1, -, A\}$$

La distance entre deux éléments est la distance euclidienne dans l'espace factoriel c'est-à-dire :

$$d^2 (i, i') = \sum_{\alpha = 1,-,A} (F_\alpha(i) - F_\alpha(i'))^2$$

ce qui correspond à la distance X^2 pour le tableau d'origine. On a, en reprenant les mêmes notations que précédemment :

$$\sum_{\alpha = 1,-,A} (F \quad (i) - F \quad (i'))^2 = \sum_{j \in J} A/f_j \ (f_j^i - f_j^{i'})^2$$

avec A suffisamment grand.

2. <u>Critère d'agrégation</u>

La classification hiérarchique est un ensemble de parties de l'ensemble I (ensemble des éléments à classifier) partiellement ordonné par inclusion ; on note C_I la classification hiérarchique.

Au niveau le plus bas C_0 de la hiérarchie sont placées les classes terminales ou terminaux de la hiérarchie.

Les autres classes de C_I sont appelées noeuds de la classification ; tout noeud est une réunion des classes qui sont placées en-dessous de lui.

Le critère d'agrégation choisi est celui du moment centré d'ordre 2 d'une partition.

Le moment d'inertie de l'ensemble I à classifier est :

$$M^2 (I) = \sum_{i \in I} p_i \, ||i - G||^2$$

avec p_i le poids de l'élément i

et G le centre de gravité.

Soient deux partitions successives dans une représentation hiérarchique :

$som_{C_{h-1}}$ et som_{C_h} (som_{C_h} = ensemble des sommets au niveau h)

<u>On observe</u> :

$$som_{C_h} = som_{C_{h-1}} \cup (a_h) - (s_h - s'_h)$$

avec a_h la nouvelle classe créée avec s_h et s'_h

$$C_h = C_{h-1} \cup (a_h)$$

avec pour conséquence :

$$M^2 (I) = M^2 (\text{som } C_h) + (M^2 (t) \; ; \; t \in \text{som } C_h)$$

et

$$M^2 (I) = M^2 (\text{som } C_{h-1}) + (M^2 (t) \; ; \; t \in \text{som } C_h - 1)$$

ainsi

$$M^2 (\text{som } C_h) = M^2 (\text{som } C_{h-1}) + M^2 (s\ h) + M^2 (s'\ h) - M^2 (s\ h \cup s'\ h)$$

d'où une formule de récurrence pour les moments centrés d'ordre 2 des partitions successives. La procédure d'agrégation consiste à agréger 2 classes s_h et s'_h de som C_{h-1} telle que M^2 (som C_h) soit maximal.

Et d'où le critère d'agrégation :

$$S (s_h, s'_h) = M^2 (\text{som } C_{h-1}) - M^2 (\text{som } C_h)$$

c'est-à-dire :

$$S (s_h, s'_h) = m_{sh}\ m_{s'h}/(m_{sh} + m_{s'h}) \ || s_h - s'_h ||^2$$

avec m_{sh} = poids de la classe s_h

$m_{s'h}$ = poids de la classe s'_h

Par conséquent, la valeur de l'incidende de niveau (en reprenant les notations du paragraphe précédent) est la suivante :

$$\nu (m) = (f_a + f_b)^{-1} f_a\ f_b \ || F (a) - F (b) ||^2 \quad \forall\ n \in \text{Nod}$$

avec Nod = ensemble des noeuds.

$$f_a = \sum_{i \in a} f_i$$

$$f_b = \sum_{i \in b} f_i$$

$$F_\alpha (a) = \sum_{i \in a} f_a^{-1} f_i\ F_\alpha (i)$$

D. Logiciels nécessaires pour utiliser les méthodes d'analyses des données

Compte tenu de l'importance des calculs nécessaires pour utiliser ces méthodes, il faut disposer d'un ordinateur et d'un ensemble de programmes adaptés.

Dans presque tous les centres de calcul, il existe des bibliothèques de programmes qui permettent :

. la saisie des données et leur vérification,

. le traitement des données par la statistique descriptive et l'utilisation des tests paramétriques (ou non paramétriques),

. de calculer une analyse multidimensionnelle, qui peut être l'analyse en composante principale, l'analyse factorielle des correspondances ou l'une des méthodes de classification automatique.

On peut citer quelques logiciels bien connus :

. les programmes de statistique "classique" (LEBART et FENELON 1975 ; LEBART et al., 1979) ;

. les programmes généraux d'analyse factorielle des correspondances (ESCOFFIER-CORDIER, 1965 ; BENZECRI, 1979 ; LEBART et al., 1979) ;

. les programmes de classification automatique pour l'analyse des données (JAMBU et LEBEAUX, 1978).

Ces programmes ont pour inconvénients :

. d'être "gros demandeurs" de mémoire centrale donc utilisables par les seuls calculateurs de grande capacité,

. d'être orientés vers le travail par lot alors que l'informatique actuelle évolue davantage vers l'utilisation en conversationnel,

. de ne pas être intégrés dans des chaînes complètes de traitement.

Il se pose par conséquent des problèmes d'interface de fichiers entre les différents traitements à exécuter.

C'est pourquoi il est préférable d'utiliser des chaînes de logiciels qui bien que moins performantes sur le plan de la vitesse d'exécution peuvent être utilisées sur des ordinateurs de petite capacité en conversationnel. En effet, très souvent, le résultat d'un premier travail détermine le choix du travail suivant.

◊ Description sommaine d'une chaîne complète de logiciel de traitement des données

Le laboratoire de statistique de l'INSERM U 49 de Rennes* a récemment développé une chaîne de programmes qui comprend :

. un programme de saisie et de manipulation des données appelé BIDOUI (JOSSE, 1982).

Très souvent, après avoir saisi des informations, il est indispensable de leur faire subir des transformations dans un but d'homogénéisation ou de calculs divers comme celui d'index synthétiques.

---

* Unité INSERM U 49 - Responsable de la section mathématique-informatique : Monsieur KERBAOL.

D'autre part, ce programme propose un grand nombre de tests de statistique descriptive unifactorielle (analyse de variances, tests paramétriques ou non, régression et corrélation). Un utilisateur dispose à tout moment de l'ensemble des fonctions et il a toute latitude dans le choix ;

. un programme de calcul et d'édition des facteurs et des contributions d'une analyse factorielle des correspondances ou d'une analyse en composantes principales (KERBAOL et ROGER, 1975) ;

. un programme d'édition des graphiques plan par plan qui permet de choisir le type d'éléments à éditer :

- variables analysées
- variables supplémentaires
- observations analysées
- observations supplémentaires
- sous-ensembles d'observations

. un programme de classification automatique des éléments non limités en nombre d'éléments à classer (BUARD, 1980). Cette classification est utilisée sur les facteurs de l'analyse des correspondances ;

. un programme d'aide à l'interprétation de l'analyse qui visualise sur écran graphique et table traçante les graphiques, calcule des ellipses de garde, visualise tous les sous-ensembles d'individus à la demande de l'utilisateur (QUENNEHEN, 1977).

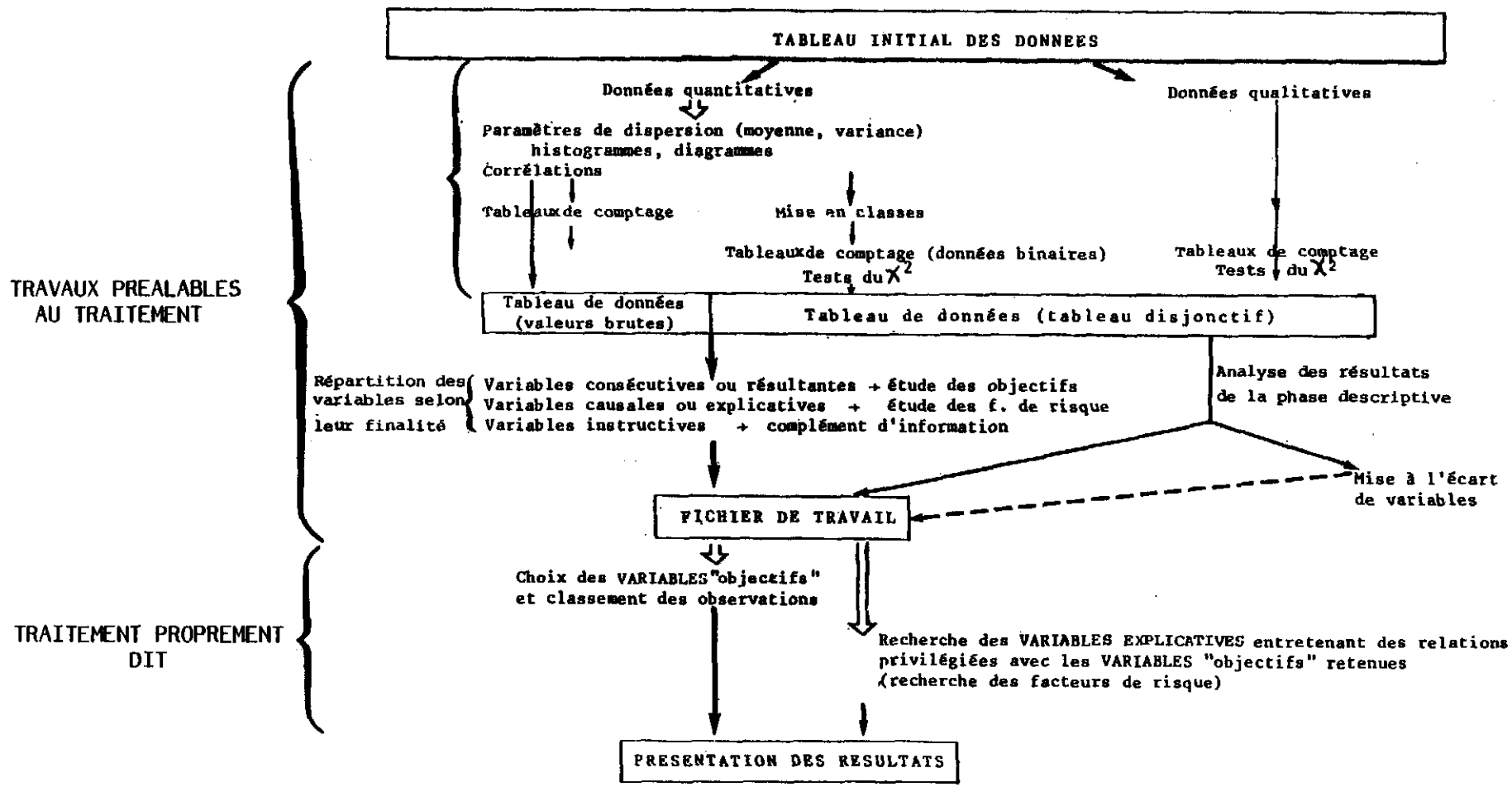
Les différents maillons de cette chaîne de programmes sont exécutables sur un ordinateur de 32 K mots de 16 bits. On peut utiliser un fichier de 500 variables brutes et calculer une analyse factorielle de 186 variables. Tous les modules sont exécutables les uns après les autres sans aucune manipulation car tous les fichiers sont compatibles entre eux. Enfin, tous les traitements s'effectuent en conversationnel, ce qui permet de choisir une méthode de traitement en fonction des résultats antérieurs.

- Conclusion -

Le traitement statistique des tableaux de données comme ceux qui proviennent des enquêtes écopathologiques requiert des méthodes adaptées aux particularités de ces dernières. Les méthodes de l'analyse des données semblent appropriées. Mathématiquement éprouvées, elles apportent toute la rigueur scientifique nécessaire à la démonstration. Elles présentent également l'avantage de constituer un bon support pour une pédagogie en raison d'une présentation des résultats facilement accessible pour un public de non spécialistes. Pour la bonne conduite des opérations il est cependant nécessaire de disposer non seulement des logiciels de calcul mais également d'aide à l'interprétation. Au cours de la seconde partie de ce rapport est exposée la démarche suivie lors du dépouillement des enquêtes.

***

Figure 1 : Organigramme des principales opérations de dépouillement d'une enquête épidémiologique concernant les maladies d'élevage chez le porc.



SECONDE PARTIE :

APPLICATION AU DEPOUILLEMENT D'UNE ETUDE ECOPATHOLOGIQUE

Au cours de la première partie ont été rapidement exposées les méthodes de l'analyse de données et les raisons de leur utilisation pour aborder l'étude des maladies d'élevage. La seconde partie traite plus concrètement de la démarche suivie pour le dépouillement d'une enquête épidémiologique. Les limites et les contraintes de la méthode sont également soulignées.

I. CONDUITE DE L'ETUDE DU TABLEAU DE DONNEES

La mise en place de l'enquête a été précédée d'une préparation très soignée du protocole, notamment l'établissement d'un tableau exhaustif et cohérent de données vérifiées. Une nouvelle vérification de vraisemblance est effectuée avant les opéraitons de dépouillement. La figure n° 1 présente l'organigramme de ces opérations successives.

1. Les travaux préalables

Ils sont relativement nombreux, difficiles et fastidieux mais impératifs. Ils comprennent successivement (figure 1) :

a. L'étude descriptive

Elle est nécessaire pour plusieurs raisons :

. Elle permet de mesurer l'étendue de la variation pour chacune des variables quantitatives. L'étude comparative n'est virtuellement possible que lorsqu'il existe une certaine dispersion dans les valeurs enregistrées.

. Outre la connaissance de l'étendue de la variation, elle permet l'appréciation de l'allure de la distribution des valeurs. Celle-ci interviendra au moment du découpage des variables en classes.

. Pour les variables qualitatives (réponse par oui ou par non par exemple), des tableaux de comptage (tableaux de BURT) permettent rapidement et simplement de connaître les effectifs au croisement des lignes et des colonnes.

. L'étude détaillée des tableaux permet également de révéler le sens de l'évolution des variables, prises deux à deux. L'examen est complété par des séries de tests simples (X^2). En fait, les renseignements que sont en mesure de fournir ces tableaux sont des plus complets, les éventuelles constructions souvent arbitraires du maître d'oeuvre étant éliminées.

. Enfin, pour les variables quantitatives, les tableaux de corrélation donnent une autre indication sur le sens de leur évolution.

b. La mise en classe des variables quantitatives

La présence simultanée de données quantitatives continues ou discrètes et des données qualitatives au sein du tableau de leur traitement simultané, entraîne l'obligation du découpage en classes des variables quantitatives. Le nombre de classes et les bornes retenues sont essentiellement fonction de l'allure de la distribution. On s'attache cependant à équilibrer au mieux les effectifs dans chacune des classes bien que cela ne soit pas toujours possible. Il existe en effet pour certaines variables des valeurs "seuils" qui ont une signification particulière (seuil de positivité d'un résultat sérologique par exemple) et doivent être retenues en priorité comme bornes en dépit d'effectifs par classe sensiblement déséquilibrés.

Il faut également ajouter que la fixation des limites de classe n'est pas obligatoirement définitive à ce stade. A mesure de l'avancement des travaux d'analyse, le nombre des classes et des limites peut être modifié pour certaines variables. Ceci s'explique par la détection de nouvelles valeurs seuils, conséquence des relations qu'entretiennent ces variables avec les autres variables analysées, ces relations n'apparaissant pas au grand jour avant l'engagement du traitement.

Quoiqu 'il en soit, à l'issue de la mise en classes, l'ensemble du tableau de données revêt une allure disjonctive complète (tableau I).

Tableau I : Passage à un tableau disjonctif complet.
Exemple d'une variable.

Nom de la variable : poids du porcelet à la naissance (kg) - symbole PMN

Dispersion des observations : mini 0,8 kg, maxi 2,1
moyenne = 1,3 σ^2 = 1,2

PMN n°	Bornes de classes (kg)	Fréquence relative
1	de 0,8 à 1	15
2	de 1,01 à 1,2	22
3	de 1,201 à 1,4	22
4	de 1,401 à 1,6	24
5	+ de 1,6	17

Il s'en suit que la ligne représentant chacun des porcelets, unité d'observation, apparaîtra sous la forme suivante :

	PMN 1	PMN 2	PMN 3	PMN 4	PMN 5
Porcelet de 1,15 kg	0	1	0	0	0
Porcelet de 1,7 kg	0	0	0	0	1
Etc.					

c. Répartition des variables selon leur finalité

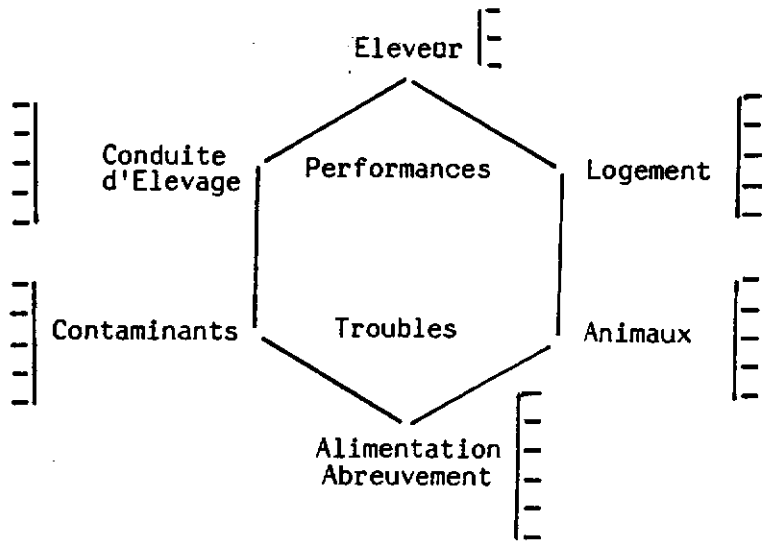
Cette phase du travail est délicate et elle montre à quel point ce type de démarche requiert une bonne coordination entre les intervenants.

La pratique des visites approfondies dans un grand nombre d'élevages a permis de classer les renseignements collectés selon un schéma hexagonal (figure 2) (TILLON et al., 1980). Le phénomène pathologique envisagé est toujours considéré comme le résultat d'une anomalie à un ou plusieurs sommets de l'hexagone. Les variables qui vont le qualifier correspondent ainsi à des variables résultantes ou consécutives. Lorsqu'on étudie les maladies d'élevage, les mortalités, symptômes, lésions et performances sont généralement rangées dans cette catégorie.

En revanche, il existe une multitude d'éléments susceptibles d'agir seuls ou associés sur la mise en route du processus pathologique. Lorsqu'en aucun cas ils ne peuvent être la conséquence du phénomène étudié, de telles variables sont dites "causales" ou "explicatives". Le traitement analytique devra en extraire les plus prépondérantes qui seront appelées plus tard des "facteurs de risque".

Enfin, il existe une autre catégorie de variables qui peuvent être à la fois causales ou résultantes. Ces variables généralement peu nombreuses lorsque la chronologie des faits est bien établie, sont dites instructives car elles correspondent à des compléments d'information et servent à discuter le résultat.

Figure 2 : Représentation de l'écosystème par une figure hexagonale.



Pour illustrer la notion de variable instructive, on peut prendre un exemple tiré de l'étude de la pathologie digestive du sevrage chez le porcelet (1979).

Il est établi que les troubles surviennent préférentiellement au cours des 20 premiers jours suivant celui du sevrage lorsque l'ingestion alimentaire dépasse les capacités de digestion du porcelet.

Une des variables (quantité d'aliment consommée par le porcelet sur cette période), en abrégé (QCPS), pourrait donc a priori être considérée comme une "cause" possible des problèmes. Il serait alors logique de la ranger dans les variables explicatives desquelles sont sélectionnés plus tard les facteurs de risque.

Il est cependant également établi que, lorsque les désordres intestinaux apparaissent, ils entraînent dans leur sillage une réduction de la consommation alimentaire. Dans ce cas, la variable (QCPS) est un résultat, une conséquence.

Ainsi, lorsque le phénomène est envisagé dans le temps, cette variable est soit explicative des troubles étudiés, soit consécutive ou expliquée par eux. Notre protocole de collecte des données ne permettant pas de segmenter, par individu, les étapes de la mise en place du processus, la variable (QCPS) a donc été considérée comme complément d'information. Elle a néanmoins participé à l'instruction du dossier d'où son nom de variable instructive.

d. La réduction du tableau initial en une base ou fichier de travail

En raison de la grande taille du fichier global qui alourdit les manipulations et entraîne une confusion, il est souvent judicieux de procéder à sa réduction en une base de travail dont le nombre de variables est limité à quelques dizaines.

◊ Première phase

La nécessaire recherche de l'exhaustivité et de la fiabilité dans la collecte conduit à certaines doublures dans un but de vérification. Il appartient au maître d'oeuvre de réaliser la sélection.

De plus, l'existence de grilles d'appréciation de tel ou tel phénomène, antérieurement mises au point, permet de ne retenir dans le fichier de travail que le résultat final de leur application. Par contre, il est délicat de recourir à des index synthétiques dont on ne connaît pas suffisamment le champ d'application ou la validité, en particulier quand on se sépare délibérément des éléments de leur calcul.

◊ Obtention du fichier de travail

Une seconde série d'éliminations intervient mais, à la différence de la première, la sélection est cette fois basée sur l'examen des tableaux de corrélation et de comptage ou encore sur le résultat de tests statistiques comme le X^2. Les variables résiduelles retenues réparties dans leur rubrique d'appartenance sur la figure hexagonale qui représente l'écosystème (figure 2) seront soumises à une première analyse factorielle des correspondances. Le plus souvent, chaque rubrique est préalablement envisagée séparément mais, pour chaque analyse concernant chacun des sommets de l'hexagone, des variables consécutives sont intégrées sous forme de variables "supplémentaires" (cf + loin). Les variables dont le déploiement se rapproche de celui des variables consécutives et ainsi considérées comme probablement "intéressantes" sont conservées dans le fichier de travail.

2. Conduite de l'analyse du tableau de données

L'analyse se déroule en deux étapes successives correspondant respectivement à la définition des objectifs puis à la recherche des facteurs de risque. La seconde étape est généralement la plus longue.

1ère étape : définition des objectifs en regard de la maladie d'élevage

Cette définition doit être précise. Elle peut être fondée exclusivement sur des critères de santé mais on peut également y intégrer des critères zootechniques (performances). Plusieurs variables consécutives de base sont ainsi en mesure de participer à la formulation du problème à condition toutefois qu'elles ne soient pas indépendantes et qu'elles évoluent dans le même sens. Il peut arriver en effet dans le cas de syndromes particulièrement complexes que le spécialiste, à la lecture des premiers résultats, s'aperçoive de sérieuses divergences dans le déploiement des variables qu'il voulait initialement considérer simultanément comme cible. Il doit alors reconsidérer sa définition et au besoin étudier séparément les groupes d'affinités de variables consécutives.

Lorsque plusieurs variables entrent dans la définition, on établit la carte des objectifs en appliquant le programme de l'analyse des correspondances. Ainsi, la localisation des variables et celle des observations vont-elles apparaître. Les différents plans factoriels sont prospectés et pour faciliter l'interprétation du résultat une classification hiérarchique est en outre réalisée sur les facteurs de l'analyse. La première étape pourrait s'arrêter à ce stade mais l'usage révèle judicieuse une mise en classes des observations. Ces classes correspondent à une échelle de réalisation simultanée des objectifs par les observations. Il est cependant probable que les différents plans de l'analyse ne donnent pas une succession de groupes bien délimités d'observations. C'est une autre raison pour appliquer la classification hiérarchique qui visualise plus schématiquement les groupes d'affinités de variables et d'observations en prenant en compte autant de facteurs de l'analyse que souhaité. Un exemple tiré de l'une de nos récentes études sur la pathologie respiratoire* illustre ces propos (figure 3).

2ème étape : la recherche des facteurs de risque

Seules les variables réputées causales ou explicatives entrent dans le champ des investigations. Deux grandes voies d'approche sont proposées, comme dans le cas de la régression :

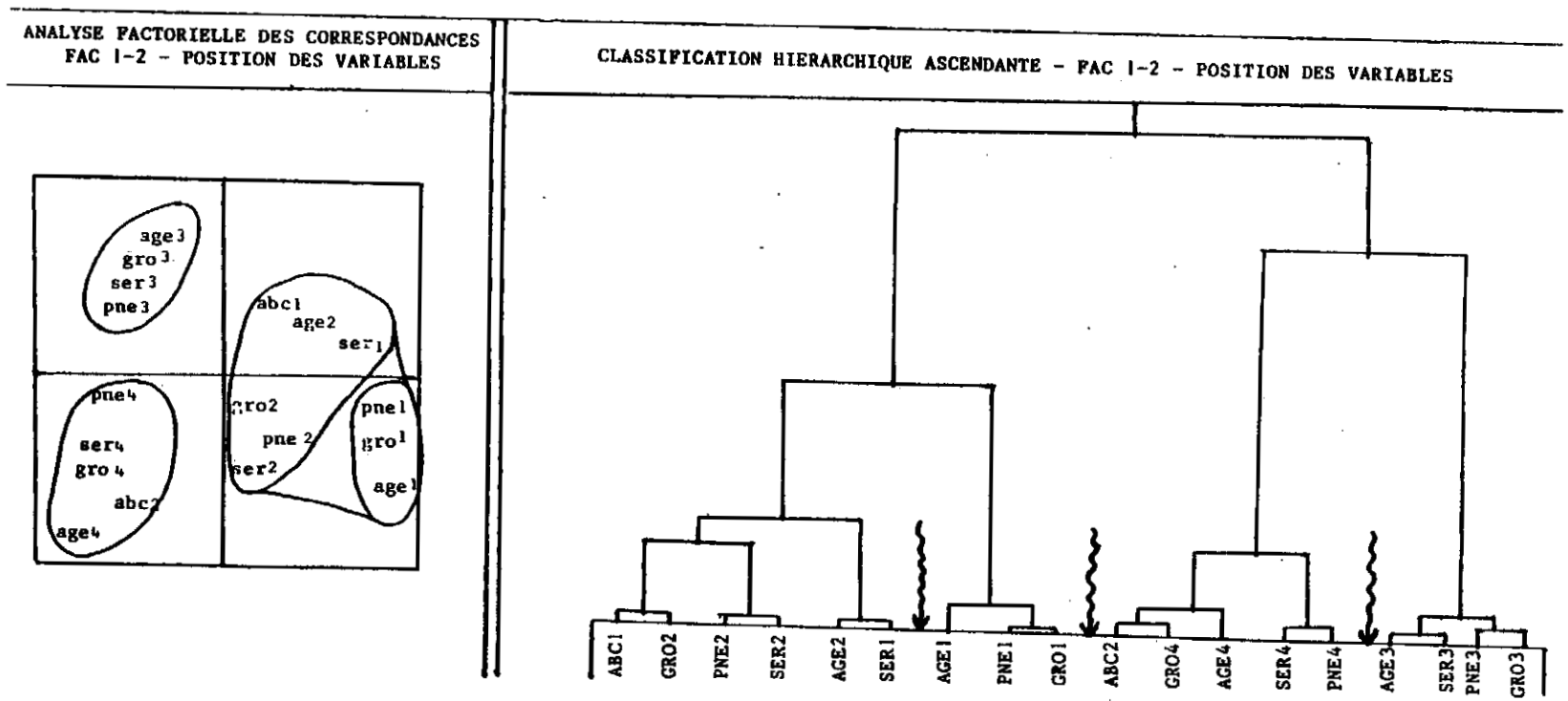
- La méthode ascendante : partant de quelques variables, une série d'analyse est engagée avec un nombre croissant de variables ;

- la méthode descendante : l'analyse de départ est plus fournie puisqu'elle comporte la totalité des variables explicatives du fichier de travail. Par la suite, on procède à des éliminations successives de variables d'influence mineure.

---

* Application d'une méthode d'étude sanitaire globale à la prévention des troubles respiratoires chez le porc à l'engrais. MADEC et JOSSE, 1981. Document Station de Pathologie Porcine - 22440 PLOUFRAGAN.

Figure 3 : Utilisation conjointe de deux méthodes d'analyse des données pour l'étude de la pathologie respiratoire en porcherie d'engraissement.



COMMENTAIRE : *La présentation "discontinue" de la classification hiérarchique (à droite) permet de constituer les groupes d'affinités de variables. Ces groupes sont alors reconstitués sur le graphique de l'A.F.C. (à gauche). Le même procédé est utilisé pour le classement des observations selon la sévérité des troubles.*

En réalité, ces considérations restent théoriques et la pratique conduit généralement à l'utilisation des deux méthodes. Une bonne connaissance du dossier à partir de l'examen détaillé des corrélations et des tableaux de comptage permet l'établissement d'une carte de départ composée de quelques variables à bonne contribution. Cette contribution est mesurée par la comparaison de leur comportement dans l'analyse avec celui des variables "objectifs". Avec la mise en place d'un "fond de carte", le comportement de nouvelles variables est étudié. Certaines, dont l'influence est favorable sur la distribution des points sont conservées, les autres sont provisoirement écartées. La base de travail reste un ensemble mouvant au point que certaines variables initialement retenues doivent être abandonnées au bénéfice d'une ou de plusieurs autres et réciproquement. La retenue ou l'abandon des variables sont fondés sur la distribution des observations induite par cette incorporation ou par ce rejet. La décision est basée sur une comparaison avec la répartition obtenue sur les variables objectifs à l'issue de la première étape. La comparaison est en effet possible puisque chaque observation conserve son identité. En outre, la constitution à partir des observations de groupes d'affinités permet l'attribution d'une symbolique pour chacun d'eux. Plus les groupes "éloignés" sur la carte des objectifs sont de nouveaux éloignés l'un de l'autre sur la carte des variables explicatives et plus le chemin parcouru dans la recherche des facteurs de risque est important.

La séquence des analyse peut être longue, mais on peut limiter raisonnablement leur nombre :

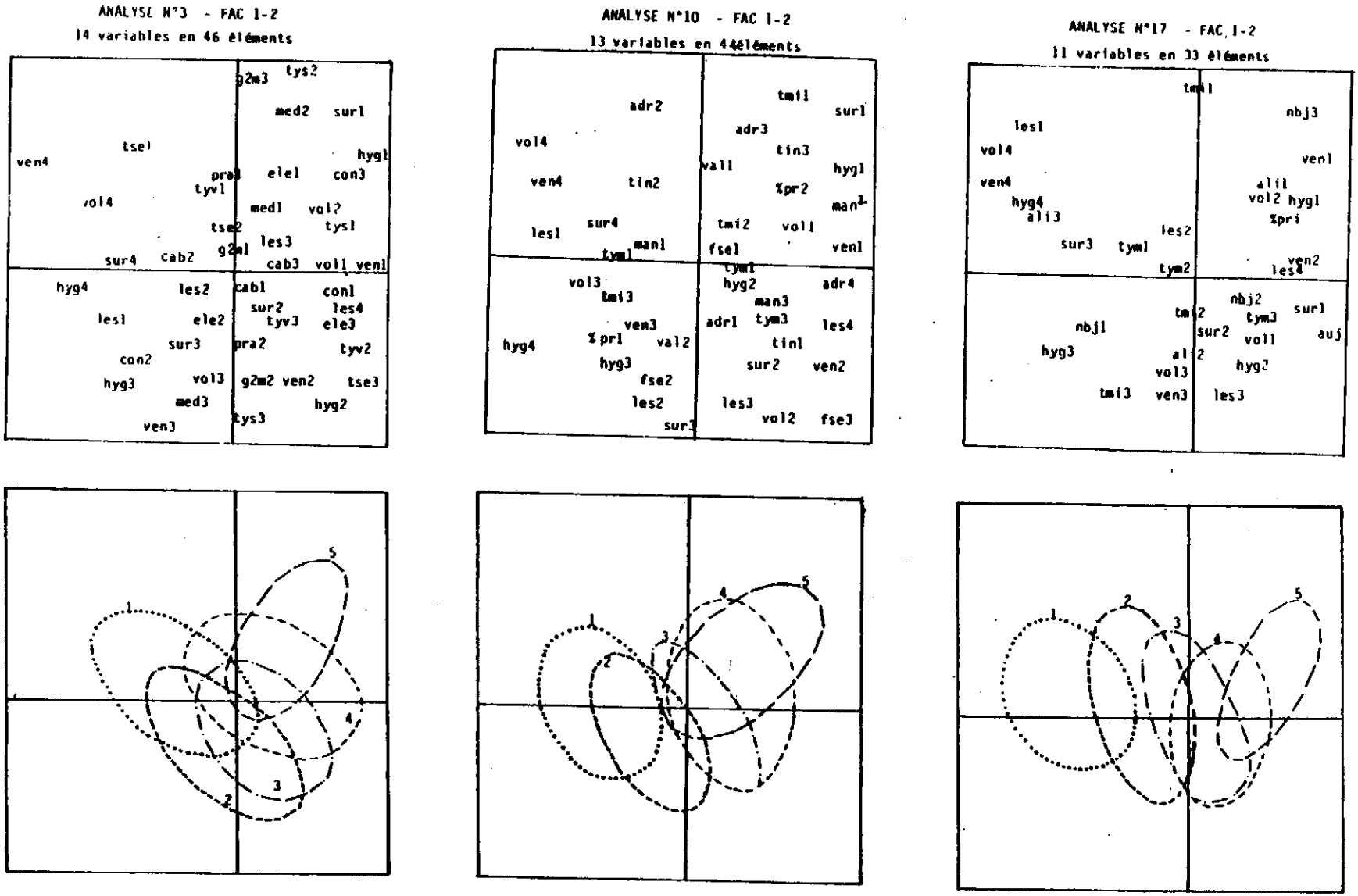
. en approfondissant les travaux de description préalable ;

. en mettant en place une organisation rigoureuse. La tenue d'un véritable plan de route est nécessaire afin de récapituler clairement les combinaisons de variables testées et les résultats obtenus. Ces données vont servir à l'élaboration des nouveaux projets d'incorporation ;

. en procédant à une étude de détail de chacune des analyses (contributions absolues et relatives des variables, étude des groupes d'affinités sur les différents facteurs...). Une ou plusieurs classifications hiérarchiques sont couplées à chaque analyse des correspondances pour aider à leur interprétation. Le comportement des variables explicatives peut ainsi être appréhendé sur les axes factoriels sur lesquels contribuent en priorité les variables "objectifs" (incorporées en qualité de variables supplémentaires - cf ci-après) ;

. en ayant recours à une technique de localisation sur le graphe de l'analyse de variables dites "supplémentaires" (FENELON, 1981). Ces dernières ne participent aucunement à la déformation du nuage mais leur situation sur la carte est cependant fonction des "relations" entretenues avec les variables analysées. Cette technique permet la présélection de variables explicatives en vue des analyses projetées. De plus, les variables "objectifs" sont ainsi introduites ce qui constitue une autre manière de visualiser le progrès réalisé au fil des analyses par la simple mesure de leur répartition sur le graphe.

Figure 4 : Présentation de trois analyses tirées d'une séquence de recherche des facteurs de risque (Etude de la pathologie respiratoire chez le porc à l'engrais). Sur les cartes du bas, seuls les "groupes objectifs" sont mentionnés, leur position est matérialisée ici par des ellipses*. On note une discrimination des groupes qui va s'accentuant de l'analyse n° 3 à l'analyse finale n° 17. Les variables qui ont permis la répartition sur l'analyse finale sont appelées "facteurs de risque".



* Ellipses de garde visualisant les groupes d'individus répondant au critère demandé (QUENEHEN, 1977).

La séquence est interrompue lorsque l'introduction ou l'abandon d'une variable ne permet plus de se rapprocher sensiblement du classement des observations obtenu au terme de la première étape. Le résultat final sera dans ces conditions d'autant plus satisfaisant que les deux répartitions seront proches.

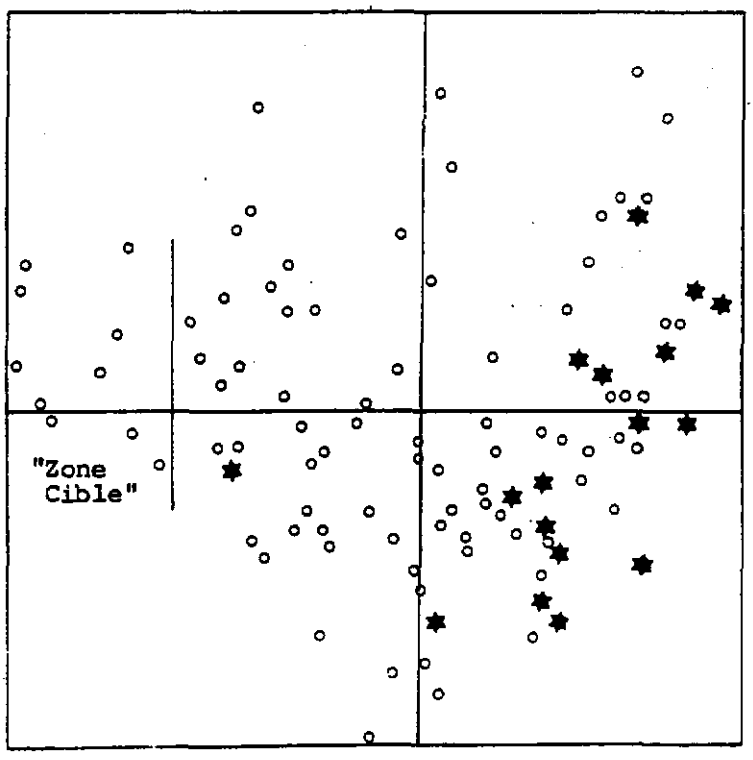
Sur la figure 4 est reporté un exemple de progression tiré de l'étude des troubles respiratoires.

TRAVAUX COMPLEMENTAIRES

. Dans le but d'illustrer le résultat obtenu, il peut être intéressant de localiser sur la carte finale les particularités de certaines observations. Ceci est possible par la procédure des critères. Ainsi, sur la figure 5 sont repérées toutes les porcheries dans lesquelles un épisode de maladie d'Aujeszky a été révélé par les contrôles sérologiques*.

Figure 5 : Localisation des élevages ayant connu un épisode de maladie d'Aujeszky chez les porcs à l'engrais (MADEC et JOSSE, 1981).

Légende : o pas de séroconversion à l'égard de la maladie d'Aujesky
✱ séroconversion



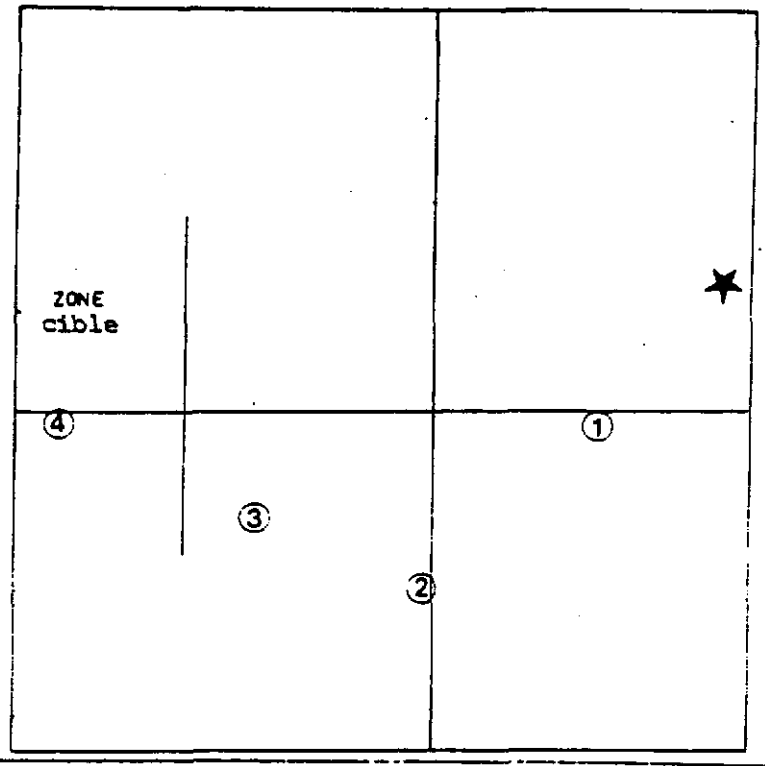
---

* Exemple de l'étude des troubles respiratoires - Réf. déjà citée.

. Il y a enfin la possibilité, dans un souci de pédagogie face à l'utilisateur du résultat, de simuler le parcours d'un élevage sur le graphique. Partant d'un profil donné, on peut, par la migration du point, apprécier l'impact théorique d'un programme de redressement fondé sur la remise aux normes de certaines variables. Sur la figure 6 est mentionné le parcours d'un élevage au départ particulièrement marqué par la pathologie respiratoire. Lorsque l'élevage se trouve en position 3 l'essentiel des troubles a disparu. On peut néanmoins poursuivre les améliorations en vue d'atteindre la zone cible dans laquelle la plupart des facteurs de risque sont maîtrisés.

Cette opération de simulation suppose la stabilité du nuage. On la vérifie par l'élimination au hasard d'une fraction des observations, ce qui ne doit pas modifier sensiblement la position des variables.

Figure 6 : Simulation de l'évolution de la localisation d'un élevage en fonction de l'amélioration des points faibles du profil.



	VENT.	THIN.	NBJA	VOL.	TYPO.	LES.	% PRIM.	ALIM.	ALU.	SURF.	HYG.	Symbole ou Numéro
Profil initial	1	2	3	2	3	3	0	1	1	1	1	★
Intervention sur la t° et la ventilation	3	3	1	"	"	"	"	"	"	"	"	1
Action sur l'hygiène, la surface et les types morphologiques	"	"	"	"	2	"	"	"	"	2	3	2
Intervention sur la maladie d'Aujeszky, sur l'alimentation et sur la qualité sanitaire du porcelet	"	"	"	"	"	2	"	3	0	"	"	3
profil cible												4

II. INTERET ET CONTRAINTES DE LA DEMARCHE

. La démarche suivie pour l'étude des maladies d'élevage est avant tout analytique et comparative. Elle prend en compte des situations réelles rencontrées dans les élevages et elle se distingue en cela de l'expérimentation dans laquelle le contrôle ou la neutralisation de la quasi-totalité des variables ne faisant pas l'objet du test expérimental est possible. Une moindre incidence des troubles par le simple fait d'exiger de l'éleveur certains enregistrements peut néanmoins être constatée (effet de HAWTHORNE). Ce phénomène n'est cependant pas nuisible dès lors que les différentes mesures sont dans tous les cas correctement réalisées. Il faut noter toutefois que pour certaines études, une modification du comportement de l'éleveur au vu de ses propres enregistrements peut gêner l'appréhension du problème. Il peut en résulter une augmentation du nombre de variables instructives* et surtout une réduction de la variabilité pour certaines variables causales ou résultantes. Or, il est fondamental de disposer d'un grand étalement des valeurs pour les différentes variables.

. Même si elle permet parfois la mise en relief de relations physiopathologiques, l'étude écopathologique comparative n'a pas pour objectif la connaissance précise des mécanismes de l'intervention des agents étiologiques. Elle se cantonne à l'étude des relations entretenues entre les animaux et leur milieu. C'est la raison pour laquelle on peut parler d'épidémiologie des circonstances. Elle est complémentaire des démarches étiologique et physiologique.

Les relations de dépendance entre les différentes variables de l'écosystème sont méthodiquement prospectées. Ainsi, toute interprétation sur l'intervention particulière de l'un ou l'autre des facteurs de risque dépistés n'a de sens que si elle est envisagée dans le cadre de l'ensemble (GRUNDY, 1973). Plus les facteurs de risque sont nombreux dans des positions défavorables et plus la probabilité de voir apparaître la maladie est élevée.

. Les facteurs de risque impliqués, pris isolément, sont rarement des découvertes qui suscitent un grand effet de surprise par leur nouveauté. A cet égard, on est même tenté de penser à l'évidence de leur implication dans le phénomène pathologique étudié. Cette réflexion parait bien décevante au terme d'une longue entreprise de collecte et de dépouillement de données. La réalité est cependant différente en raison du grand nombre de variable considérées pour chaque étude qui entraîne un très grand nombre de combinaisons possibles, dont plusieurs peuvent apparaître a priori évidente pour le lecteur. Seule l'application de la démarche permet la détection objective de la combinaison la plus pertinente (FENELON, 1981). C'est un avantage essentiel de la méthodologie biométrique décrite par ailleurs dans d'autres domaines (MIETTINEN, 1976).

---

* On peut prendre un exemple dans l'étude de la pathologie de la mise-bas chez la truie (MADEC et JOSSE, 1982). L'observation de températures rectales élevées en dehors de toute manifestation clinique incite certains éleveurs à procéder aux injections d'antibiotiques. Ces interventions n'auraient probablement pas été réalisées sans la connaissance des température rectales.

Pour illustrer ces propos on peut citer ici les réponses à une interrogation écrite proposée à la Station de Pathologie Porcine de Ploufragan aux binômes "enquêteur-enquêté" avant publication des résultats et relative à la pathologie respiratoire sur laquelle ils venaient de travailler.

La question était la suivante : "Quelles sont d'après l'expérience que vous avez de l'élevage de porcs, les causes du développement des maladies respiratoires et des mauvaises croissances ?"

On obtient les résultats suivants :

. nombre de bulletins rédigés 154

. nombre total de causes incriminées 72

. nombre de bulletins comportant 10 ou les 11 facteurs de risque 0

. nombre de bulletins comportant plus de 7 des 11 facteurs de risque . . . . . 16 (10 %)

. nombre de bulletins comportant moins de 5 des 11 facteurs de risque . . . . . 63 (41 %)

Sur les 11 premières variables citées par l'ensemble du groupe par ordre de fréquence, seules 6 sont des facteurs de risque révélés par les méthodes de l'analyse des données ! Non seulement l'évidence de chacun des facteurs de risque est loin d'être acquise mais la "bonne" combinaison n'a jamais été trouvée.

. Les résultats de l'étude sont directement utilisables par les praticiens de l'élevage.

En présence d'un cas pathologique, la connaissance de la situation à l'égard des facteurs de risque puis la mise en place d'un programme concerté de redressement des éléments déficients doit permettre la régression des manifestations cliniques. Les facteurs de risque réputés prépondérants pour l'ensemble de l'échantillon ne sont pas cependant obligatoirement les seuls en cause lorsqu'on envisage chacun des élevages du groupe. Dès lors, l'utilisateur doit envisager en priorité la liste des facteurs de risque mais il ne doit pas négliger pour autant d'autres paramètres particuliers à chaque situation et pour lesquels les valeurs s'écartent notablement des recommandations.

Le champ d'application des résultats de ces études ne se limite pas aux seuls élevages très marqués par la pathologie. En présence d'une situation apparemment convenable, la mesure des facteurs de risque donne une appréciation sur la probabilité d'apparition prochaine des toubles. Une action de prévention peut être entreprise lorsque des variables sont défaillantes bien que cela ne se traduise pas encore par des signes cliniques perceptibles.

. Les résultats obtenus ne peuvent pas avoir de valeur définitive ; ils sont susceptibles de devenir caduques en raison du caractère très changeant de certaines variables explicatives et de certains facteurs de risque.

Une actualisation périodique de la carte des facteurs de risque est indispensable pour tenir compte de la fluctuation temporelle et pour assurer à l'utilisateur la pertinence et la véracité de son contenu. L'efficacité de l'utilisation des résultats à l'extérieur de la zone géographique de première collecte ne peut pas, par contre, être toujours assurée. En effet, il y a de fréquentes divergences régionales dans les modes d'élevage et les conditions de milieu (densité d'animaux, habitat, climat...). Toutefois, les derniers travaux à propos de la pathologie digestive du sevrage démontrent qu'ils peuvent également servir de canevas dans les autres régions de notre pays (MADEC et al., 1982).

. La mise en oeuvre de la méthode est onéreuse, contraignante et l'obtention d'un résultat probant reste incertaine si l'exécution du travail n'est pas rigoureuse. L'examen du tableau de données peut en effet ne pas aboutir à la sélection des variables qui offrent une répartition satisfaisante des observations. Une telle issue peut avoir diverses origines parmi lesquelles :

- une mauvaise formulation des objectifs,
- le manque d'exhaustivité du tableau initial de données,
- la mauvaise qualité de l'information (valeurs trop approximatives),
- la prospection insuffisante des combinaisons de variables,
- l'absence de liaisons suffisamment fortes entre les variables explicatives. Dans ce dernier cas, on s'éloigne du domaine des maladies déterminées par l'intervention cumulative de plusieurs facteurs.

CENTRAL DE RECHERCHES BIBLIOTHÈQUE

. De telles enquêtes supposent l'harmonisation des méthodes de travail des divers intervenants notamment des laboratoires. De plus, la complexité des sujets traités nécessite un effectif suffisant d'observations*. Le groupe de base peut alors être subdivisé et les comparaisons se poursuivre entre les observations à l'intérieur des sous-groupes.

Par ailleurs, cette méthode d'étude est nécessairement multidisciplinaire car le champ des compétences requises est vaste : pathologie, zootechnie, génie rural, statistique, informatique. Enfin, aux moyens humains pour la collecte, la saisie et le traitement des données, il y a lieu de joindre les moyens de calcul appropriés (ordinateur et logiciels).

- Conclusion -

Les méthodes d'analyse des données permettent d'aborder l'étude épidémiologique des maladies d'élevage qui représentent aujourd'hui la plus grande part des pertes économiques dans certaines espèces comme le porc. Leur mise en oeuvre réclame cependant certaines conditions qu'il est parfois difficile de réunir. L'intérêt de ces méthodes réside tant dans la rigueur de leur fondement scientifique que dans la présentation originale des résultats qui en sont issus.

---

* Une centaine de troupeaux ont été considérés pour chacune des études réalisées à ce jour.

Les techniques de calcul, conduisent à adopter des règles nouvelles dans le dépouillement des enquêtes. Bien que la conclusion des travaux contraste avec les propos souvent fragmentaires issus de l'application de tests statistiques dans le contexte rigide de travaux expérimentaux, il demeure que les deux approches sont complémentaires pour l'étude des phénomènes pathologiques.

- Bibliographie -

1. BENZECRI J.P. (1964).- Cours de linguistique mathématique. Publication multigraphiée (Faculté des Sciences de Rennes), 38 pages.

2. BENZECRI J.P. (1979).- L'analyse des données :
   I. La taxinomie (640 pages)
   II. L'analyse des correspondances (616 pages)
   Dunod - Ed. Paris (2ème édition).

3. BUARD J.L. (1980).- Thèse de doctorat de troisième cycle. Université de Rennes I, 192 pages.

4. BRUYNOOGHE (1978).- Classification ascendante hiérarchique des grands ensembles de données. Cahiers de l'Analyse des Données n° 1, 7-33.

5. ESCOFFIER-CORDIER B. (1965).- Thèse de troisième cycle. Les cahiers du Bureau Universitaire de Recherche Opérationnelle en 1969.

6. ESCOFFIER B. (1979).- Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Cahiers de l'Analyse des Données, vol. IV, n° 2, 137-146.

7. FENELON J.P. (1981).- Qu'est-ce que l'analyse des données ? LEFONEN éd. Paris, 311 pages.

8. GRUNDY P. F. (1973).- A rational approach to the "at risk" concept. Lancet 2, Déc., 29 pp. 1489.

9. JAMBU M. (1976).- Sur l'intervention mutuelle d'une classification hiérarchique et d'une analyse des correspondances. Rev. Stat. Appl. 24 n° 2.

10. JAMBU M., LEBEAUX M.O. (1978).- Classification automatique pour l'analyse des données. Tome I : Méthodes et Algorithmes, Tome II : Logiciels. Dunod. Ed. Paris.

11. JAMBU M. (1980).- Classification automatique pour l'analyse des données. I. Méthodes et algorithmes. Dunod Ed. Paris (2ème édition).

12. JOSSE J., KERBAOL M. (1982).- Programme de traitement et de manipulations de données pour mini-ordinateurs. Station de Pathologie Porcine (26 pages), non publié.

13. KERBAOL M., ROGER J.M. (1975).- Programme d'AFC-ACP pour mini-ordinatauers. INSERM U 49 Rennes, non publié.

14. LEBART L., FENELON J.P. (1975).- Statistique et informatique appliquées. Dunod Ed. Paris (3ème édition).

15. LEBART L., MORINEAU A., FENELON J.P. (1979).- Traitement des données statistiques. Méthodes et Programmes. Dunod Ed. Paris, 510 pages.

16. MADEC F., JOSSE J. (1981).- Application d'une méthode d'étude sanitaire globale à la prévention des troubles respiratoires chez le porc à l'engrais. Document Station de Pathologie Porcine de Ploufragan, 85 pages, non publié.

17. MADEC F., JOSSE J., CHANTAL A. (1982).- Evaluation d'une méthode multifactorielle dans l'analyse des troubles digestifs du sevrage. J. Rech. Porcine en France, 14, 379-386.

18. MIETTINEN O. (1976).- Estimability and estimation in case referent studies. Am. Epidemiol., 103, 226-235.

19. QUENNEHEN F. (1977).- Un langage conversationnel de visualisation graphique des résultats d'AFC. Rapport D.E.A. Université Rennes I - UER Math. et informatique.

20. TILLON J.P. (1980).- Epidémiologie du porc liée à l'élevage intensif. Journées de la Recherche Porcine en France, 12, 361-380, ITP Ed. Paris.

21. TILLON J.P., MEURIER C., MADEC F. (1980).- Estimation des pertes économiques en élevage porcin Naisseur-Engraisseur de type intensif. Bull. Off. Inter. Epizoot., 92, 371-385.

*
* *