Modélisation des données temporelles et rôle du graphisme

Françoise Lescourret

Vet Res (1994) 25, 140-146

Modélisation des données temporelles et rôle du graphisme

F Lescourret

INRA, centre de Clermont-Ferrand-Theix, laboratoire d'écopathologie, 63122 Saint-Genès-Champanelle, France

Résumé — L'importance du graphisme dans la modélisation des données temporelles, à laquelle l'écopathologie est souvent confrontée, est illustrée par 2 exemples. Le premier concerne des chroniques univariées (mesures de germes du lait de tank). Il utilise le corrélogramme (choix du modèle à ajuster), les graphiques des résidus du modèle (diagnostics) et l'expression graphique d'une analyse en composantes principales (comparaison des chroniques). Le deuxième exemple concerne des chroniques climatiques multivariées mensuelles. Une analyse multitableaux assistée du graphisme permet d'affecter à chaque station une note résumant son climat sur un ensemble de mois.

chronique / corrélogramme / modèles AR, MA / analyse multitableaux / graphisme

Summary — Temporal data modelling and role of graphism. *The importance of graphism for temporal data modelling, which is often used in ecopathology, is illustrated through 2 examples. The first concerns a univariate time series (tank milk germ measurements). It uses correlograms (choice of the model), plots of residuals of the model (diagnostics) and principal component analysis graphical outputs (comparison of time series). The second concerns climate multivariate monthly time series. A 3-way analysis and graphical outputs are used to give a mark to each station, describing its climate for a set of months.*

time series / correlogram / AR, MA models / 3-way data analysis / graphism

INTRODUCTION

L'écopathologie est confrontée à la modélisation des données temporelles, où l'on cherche à savoir si le temps est un élément structurant la variation des données et la nature de cette structure éventuelle, à des fins d'explication ou de prévision. Dans ce cadre, une partie des besoins concerne des chroniques : observations régulières de variables concomittantes à la maladie, par exemple paramètres climatiques ou descripteurs des conduites d'élevage.

On compte 2 types de chroniques : les chroniques univariées, et les chroniques multivariées, dont l'étude récente appartient au domaine de l'analyse des données structurées (Kiers, 1988). Dans les 2 cas, le graphisme peut aider considérablement le choix et le diagnostic des modèles (Box and Jenkins, 1970 ; Auda, 1983).

Deux exemples ont été utilisés. Le premier concerne des chroniques de mesures de germes du lait de tank dans des élevages laitiers, indicatrices d'éventuels dysfonctionnements hygiéniques ou sanitaires,

qu'il s'agit d'étudier à des fins d'explication ou de prévision, et de comparer entre elles. Le second concerne des mesures mensuelles de paramètres climatiques dans des stations, facteurs de risque potentiels des maladies en élevage ; l'objectif est d'affecter à chaque station une note résumant son type de climat sur un ensemble de mois.

MATÉRIEL ET MÉTHODES

Chroniques univariées

Quarante-cinq chroniques composées d'une trentaine de mesures régulières des germes du tank, ont été étudiées à l'aide du corrélogramme. Le corrélogramme est le schéma des corrélations entre les observations réalisées aux temps t et $t-k$ en fonction du pas de temps k, ce dernier étant le nombre d'unités de temps utilisées pour décaler la série. Il révèle les patrons de variation sous-jacents (tendances, alternances, corrélations à court terme...). Sur une série dont on a extrait la tendance et les variations périodiques, il permet de décider du meilleur modèle à ajuster : il décroît régulièrement pour un modèle autorégressif d'ordre p (AR(p)), il se coupe au pas q pour un modèle en moyennes mobiles d'ordre q (MA(q)), ... (Box et Jenkins, 1970 ; Chatfield, 1984). Une fois les modèles choisis et ajustés, le diagnostic s'appuie sur l'examen graphique des résidus, notamment sur les corrélogrammes des résidus. Nous présenterons 3 exemples d'application.

Pour être comparées, les 45 chroniques ont été réunies dans une matrice correspondant à une même fenêtre d'observation de 30 périodes, qu'on a soumise à une analyse en composantes principales suivie d'une classification hiérarchique à liens complets sur la base des distances euclidiennes calculées à partir des axes principaux.

Chroniques multivariées

Quarante chroniques climatiques relatives à des stations (station = élevage expérimental de bovins laitiers une année donnée) ont été étudiées pour une période de temps couvrant 4 «dates» (mois de juillet à octobre). Six variables mensuelles décrivent ces chroniques : les moyennes des températures journalières minimale *(tn)* ou maximale *(tx)*, la moyenne des écarts thermiques journaliers *(ec)*, le cumul des précipitations *(pl)*, le nombre de jours avec plus de 10 mm de précipitations *(jp)* ou avec tx-tn > 18°C *(je)*. On a choisi parmi les analyses multitableaux disponibles l'analyse triadique (Thioulouse et Chessel, 1987), qui peut être mise en œuvre avec un simple programme d'ACP. Cette analyse procède en 3 étapes, selon une philosophie générale exposée par ailleurs (Lavit, 1988) : *l'interstructure* des tableaux fournit une typologie des dates et, pour chaque variable, une description des stations relative à la typologie des dates ; le *compromis* est un résumé le plus proche possible des situations à chaque date ; *l'intrastructure* est la représentation des stations (ou des variables) aux diverses dates, qui permet de juger des évolutions.

Les méthodes ont été mises en œuvre grâce au logiciel S-PLUS (Statistical Sciences, 1991).

RÉSULTATS

Chroniques univariées

Le tracé des données en fonction du temps suggère l'absence de tendance ou de variations périodiques, ce que confirment les corrélogrammes (fig 1). La chronique 1 paraît aléatoire (aucune autocorrélation significativement différente de 0). Pour les 2 autres chroniques, supposées stationnaires, les corrélogrammes suggèrent d'appliquer des modèles d'ordre faible, par exemple AR(1) pour la chronique 2 et MA(1) pour la chronique 3. Concernant le diagnostic des modèles (fig 2), les corrélogrammes des résidus confirment l'hypothèse d'indépendance. Les corrélogrammes et les diagrammes des résidus réduits montrent qu'il n'y pas évidence de structure temporelle des résidus, mais présence perturbante d'individus déviants (fig 1).

Le premier plan de l'analyse conjointe des 45 chroniques (35% de l'inertie ; fig 3)

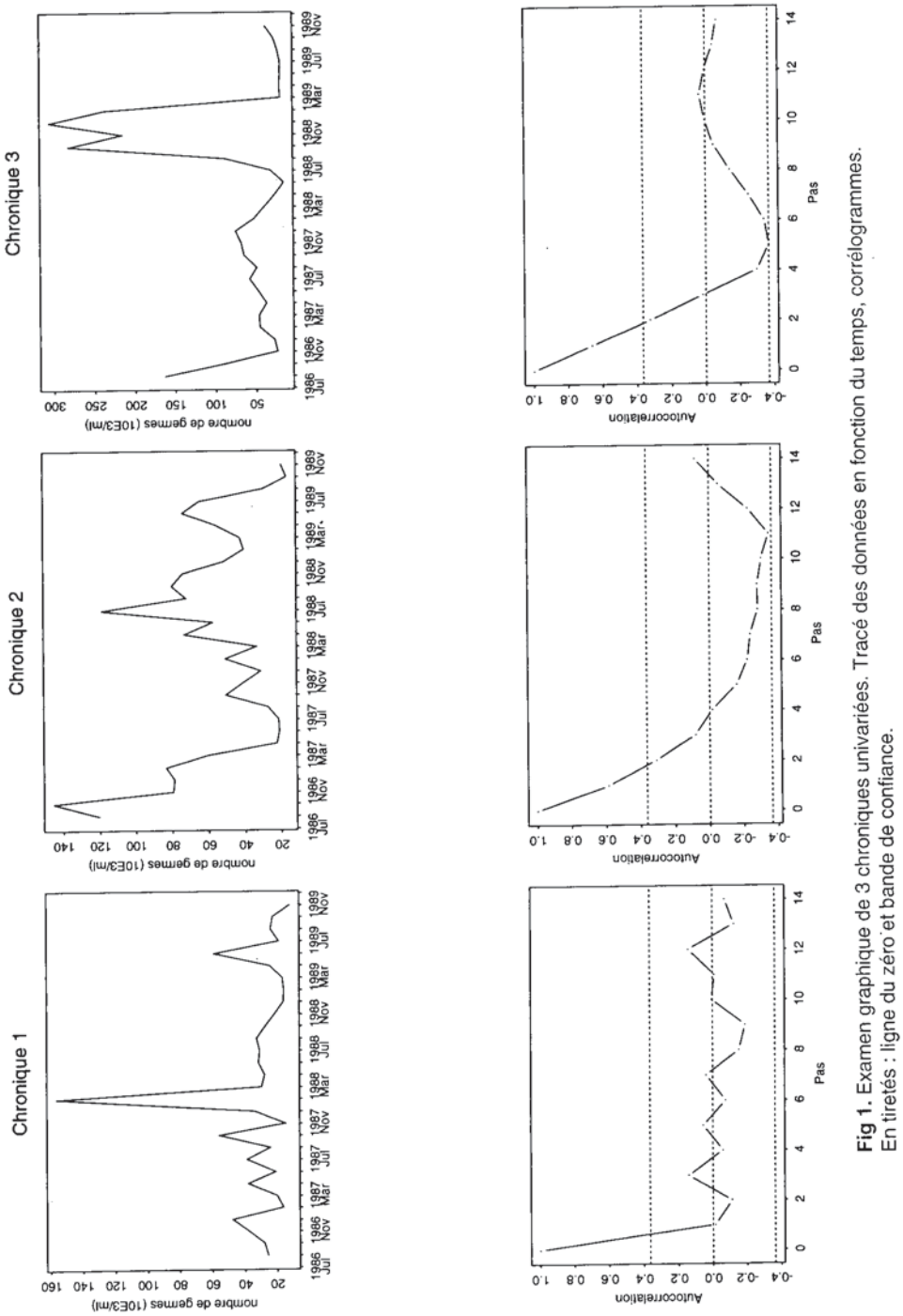

Fig 1. Examen graphique de 3 chroniques univariées. Tracé des données en fonction du temps, corrélogrammes. En tiretés : ligne du zéro et bande de confiance.

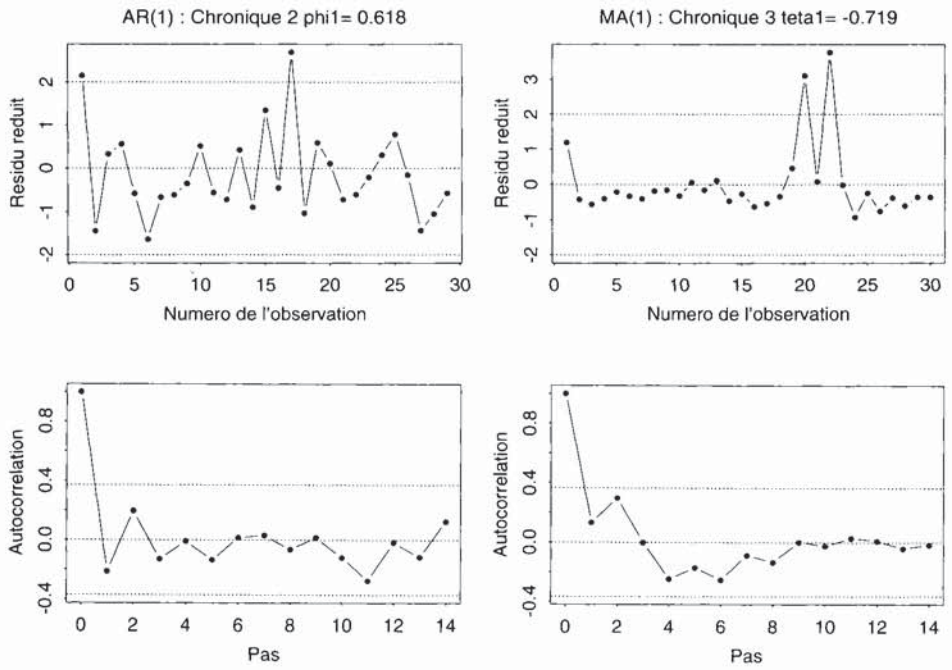

Fig 2. Application de modèles AR(1) et MA(1) à 2 chroniques univariées. Diagrammes des résidus réduits et coefficients des modèles (notations de Box et Jenkins, 1970), corrélogrammes des résidus. En tiretés : ligne du zéro et bande de confiance.

est éclairé par la classification en 5 groupes à partir de ces 2 premières composantes, et le tracé des profils moyens des groupes. Par exemple on note que la classe 4, qui réunit 26 chroniques, se situe dans le quart inférieur gauche du plan, qui correspond d'après le cercle des corrélations et le profil moyen à un niveau général faible et à une absence de pics.

Chroniques multivariées

Le premier axe de l'analyse de l'interstructure (46% de l'inertie) exprime une ressemblance entre dates, alors que le deuxième axe (23% de l'inertie) exprime la divergence entre les mois d'août et d'octobre (fig 4A). L'analyse du compromis fournit les patrons moyens sur l'ensemble des mois ; la classification des stations à partir des 6 axes de l'ACP du compromis met en évidence 4 groupes facilement caractérisables à l'aide du patron des variables (fig 4B). Par exemple le groupe 1 qui contient la station 26 est marqué par des températures journalières minimales (tn) hautes ; le groupe 2 qui contient la station 27 est marqué par des températures journalières minimales assez basses et des écarts thermiques journaliers *(ec, je)* forts ; le groupe 4 qui contient la station 30 est marqué par une pluviosité *(pl, jp)* assez élevée, une température journalière maximale *(tx)* assez basse et des écarts thermiques journaliers assez faibles.

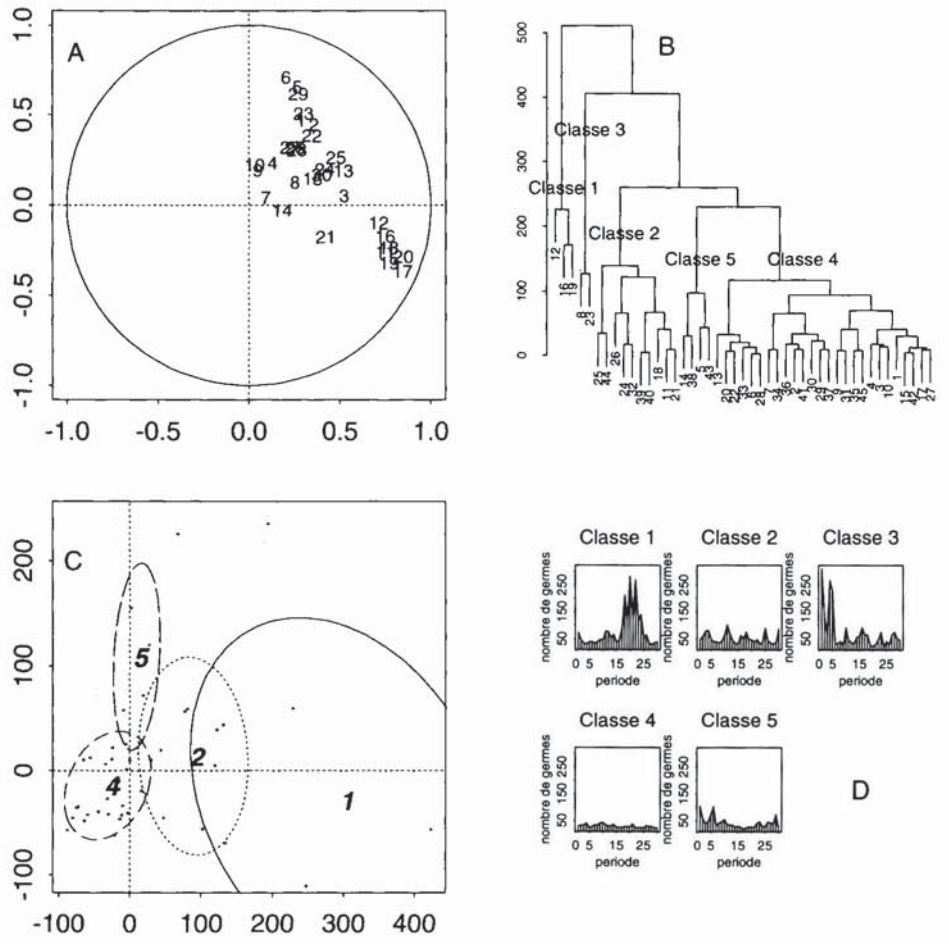

Fig 3. Comparaison de 45 chroniques univariées relatives à 30 périodes par ACP (plan F1 x F2). **A**: cercle des corrélations des variables (1 à 30) ; **B**: classification des 45 chroniques ; **C**: projection des individus avec les ellipses d'inertie des classes ; **D**: profils moyens des 5 classes.

Les numéros de classe ont été retenus comme notes résumant le climat des stations pour l'ensemble des mois. Dans l'analyse de l'intrastructure, les trajectoires d'évolution des stations (fig 4D) s'interprètent en référence au compromis (fig 4B). Par exemple, la station 26 commence avec un mois de juillet pluvieux (*cf* les positions de *pl* et *jp* dans le compromis), poursuit par un mois d'août très chaud *(tn, tx)* et termine par un mois d'octobre assez frais *(tn)*. La trajectoire de la station 27 est marquée par une réduction des écarts thermiques, une augmentation de la pluviosité et une baisse des températures régulières entre août et octobre. La trajectoire de la station 30 va dans le sens inverse pour la pluviosité et les écarts thermiques entre juillet et septembre, et s'achève par un mois d'octobre très clément *(tn, tx)*.

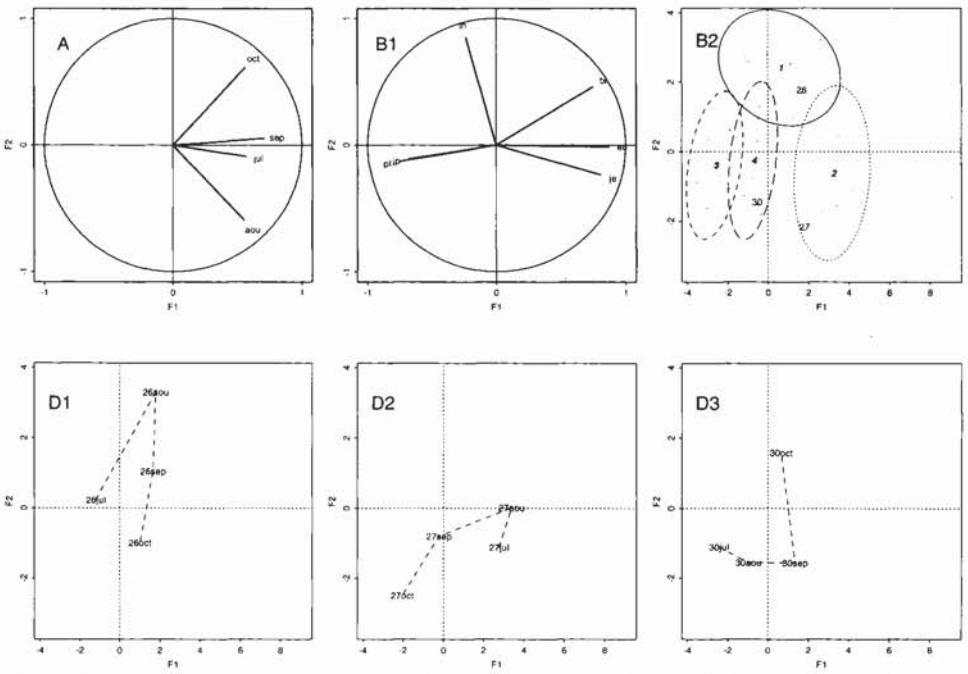

Fig 4. Analyse triadique de chroniques multivariées. Cercles des corrélations et plans F1 x F2. *Interstructure* (A : dates); *compromis* (B1 : variables, B2 : stations – points, ou numéros pour les stations 26, 27 et 30, avec les ellipses d'inertie de leurs classes) ; *intrastructure* (D1 à D3: trajectoires d'évolution des stations 26, 27 et 30).

DISCUSSION ET CONCLUSIONS

Les exemples présentés ont montré en quoi les méthodes graphiques interviennent en complément des statistiques des processus temporels. Leur utilité s'est manifestée à différentes étapes, de la simple représentation des données jusqu'à l'aide à l'interprétation des sorties des analyses statistiques (*cf* Auda, 1983). Au-delà du cadre des données temporelles, l'importance du graphisme dans le traitement de l'information, spécialement en EDA *(Exploratory Data Analysis),* n'est plus à démontrer (Mallows et Tukey, 1982); or l'EDA est une phase cruciale des recherches en écopathologie. Dans ce cadre sont proposées des stratégies interactives à base de graphisme fondées sur le pouvoir de l'œil humain à identifier des structures (Weihs et Schmidli, 1990). En outre, le graphisme est un outil de communication sans égal dans les échanges multidisciplinaires (Auda, 1983 ; Thioulouse *et al,* 1991) qui s'instaurent naturellement dans une discipline comme l'écopathologie.

RÉFÉRENCES

Auda A (1983) Rôle des méthodes graphiques en analyse des données : application au dépouillement des enquêtes écologiques. Thèse de 3e cycle, Université Lyon I

Box GEP, Jenkins GM (1970) *Time series analysis. Forecasting and control.* Holden-Day, San Francisco

Chatfield C (1984) *The analysis of time series: an introduction.* Chapman and Hall, New York

Kiers HAL (1988) Comparison of "Anglo-Saxon" and "French" three-mode methods. *Stat Anal Données* 13, 14-32

Lavit C (1988) *Analyse conjointe de tableaux quantitatifs*. Masson, Paris

Mallows CL, Tukey JW (1982) An overview of techniques of data analysis emphasizing its exploratory aspects. *In: Some recent advances in statistics* (de Oliveira JT, Epstein B, eds). Academic Press, London, 111-172

Statistical Sciences (1991) *S-PLUS User's Manual – S-PLUS Reference Manual*. Statistical Sciences, Seattle, Washington

Thioulouse J, Chessel D (1987) Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Oecol* 8, 463-480

Thioulouse J, Devillers J, Chessel D, Auda Y (1991) Graphical techniques for multidimensional data analysis. *In: Applied multivariate analysis in SAR and environmental studies* (Devillers J, Karcher W, eds). ECSC, Brussels, 153-205

Weihs C, Schmidli H (1990) OMEGA (Online Multivariate Exploratory Graphical Analysis): routine searching for structure. *Stat Sci* 5, 175-226

Vet Res (1994) 25, 146-152

Exploration de données par l'analyse post-factorielle : utilisation d'un logiciel d'interrogation de données structurées

B Faye [1]*, JM Bernard [2]

[1] INRA de Theix, laboratoire d'écopathologie, 63122 Saint-Genès-Champanelle;
[2] CNRS (URA 1201), Groupe «Mathématique et psychologie», université Paris V, 75005 Paris, France

Résumé — Le langage d'interrogation des données (LID) permet l'exploration graphique et numérique d'un nuage de points issu d'une analyse factorielle. Dans la présente communication, l'intérêt d'une analyse post-factorielle en épidémiologie est montré : mesure de l'effet département, mesure de l'effet année, évaluation de l'ajustement selon un facteur dans le cadre d'une enquête portant sur les facteurs de risque de la qualité du lait.

statistique / informatique / analyse de données / qualité du lait / épidémiologie

Summary — Data exploration by post-factorial analysis: use of a software of structured data interrogation. *The language for interrogating data (LID) allows the graphical and numerical exploration of a cloud of points resulting from a factorial analysis. The interest of a post-factorial analysis in epidemiology is shown in the present paper: a measure of the department effect; a measure of year effect; an assessment of the adjustment on one factor in the framework of a survey on the risk factors of milk quality.*

statistics / computer science / data analysis / milk quality / epidemiology

* Correspondance et tirés à part.