MODELES LINEAIRES GENERAUX ET CAPACITE DE REPRESENTATION

Francis Laloë, Nicolas Pech, Monique Simier [a]

Le recours au modèle linéaire statistique est extrêmement fréquent pour l'analyse des données dans l'étude des pêches, depuis le très classique modèle de Robson jusqu'aux régressions utilisées par exemple pour établir des relations taille-poids.

Le modèle linéaire consiste en une description de la *distribution* d'une ou plusieurs variables Y conditionnellement à des valeurs prises par des variables "explicatives", qualitatives ou quantitatives.

Le modèle linéaire est un cas particulier d'un modèle de type

$$Y = f(x, \theta) + \varepsilon$$

où $f(x, \theta)$ est l'espérance de la variable Y, une fonction des variables explicatives et de paramètres inconnus θ et où ε est une variable aléatoire centrée.

Le modèle est linéaire si $f(x, \theta)$ est une combinaison linéaire des paramètres θ.

Ainsi la régression linéaire simple $((Y_i) = x_i \times b + 1 \times a + \varepsilon_i)$ est un modèle linéaire au même titre qu'un modèle

$$E(Y_{ik}) = a_i \times x_{ik}^2 + b_i \times x_{ik} + c_i \times 1 + \varepsilon_{ik}$$

utilisé pour décrire des relations entre espérances de tailles de crevettes selon des fonctions paraboliques de la salinité, fonctions dont les paramètres diffèrent selon la vitesse d'un courant pouvant prendre 5 valeurs ($i = 1 \cdots 5$). (figure 1, données de L. Le Reste)

Les logiciels statistiques permettent de définir un tel modèle et d'en estimer les paramètres avec un nombre très réduit d'instructions. La procédure GLM dans SAS (General Linear Model) permet ainsi de réaliser cette estimation par la méthode des moindres carrés (choisir les valeurs $\widehat{\theta}$ pour des estimations de θ telles que la somme des carrés des valeurs observées et ajustées soit minimale). Cette méthode des moindres carrés présente un intérêt général ; elle est de plus particulièrement adaptée au cas où les variables Y sont indépendantes, gaussiennes et de même variance σ^2 (auquel cas les seules inconnues du problème sont θ et σ^2).

[a]HEA, Centre Orstom de Montpellier, BP 5045, 34032 Montpellier Cédex 1

La théorie du modèle linéaire est relativement ancienne. Les capacités de calcul actuelles permettent d'en tirer parti de façon de plus en plus efficace et de mettre en œuvre des généralisations pouvant être très utiles. Parmi celles-ci les modèles non paramétriques faisant appel aux méthodes d'estimation fonctionnelle sont très intéressants.

Le développement de ces méthodes ne doit cependant pas conduire à un oubli du modèle classique qui peut présenter une souplesse particulièrement efficace comme peut l'illustrer l'exemple suivant portant sur la description d'une série chronologique de températures de surface sur la côte ouest de la Côte d'Ivoire (données COADS mises à disposition par C. Roy). Les données sont des moyennes mensuelles d'observations pendant 27 années (figure 2).

Une méthode non paramétrique "STL" a été développée par Cleveland *et al* (1990). Cette méthode consiste en une description du signal périodique, et de son éventuelle modification au cours du temps. Pour ce faire, l'idée de base est d'analyser séparément à l'aide d'un lisseur (LOESS) la tendance interannuelle de chacune des douze séries de données correspondant à un mois particulier. Le résultat est alors présenté sous forme graphique avec une composante interannuelle commune à toutes les sous séries, une composante saisonnière "évolutive" et une composante résiduelle. La figure 3 présente ainsi la somme de la composante interannuelle générale et de la composante saisonnière. L'ajustement a été réalisé avec le logiciel Splus. On observe l'amplification de la baisse de température au cours de la petite saison froide, phénomène pouvant expliquer l'augmentation considérable des captures de sardinelles dans cette zone (Pézennec et Bard 1992).

En fait la démarche employée ici peut être reproduite par un modèle linéaire décrivant chacune des douze sous-séries par un polynôme de degré 5 selon le numéro du mois (codé de 1 à 324)

Ce modèle peut s'écrire sous forme :

$$Y_{ik} = a_i + b_i \times t + c_i \times t^2 + d_i \times t^3 + e_i \times t^4 + f_i \times t^5 + \varepsilon_{ik}$$

où i est le numéro du mois (de 1 à 12), k celui de l'année (1 à 27) et $t = i + 12(k-1)$.

La somme des tendances saisonnière et interannuelle issues de l'ajustement de ce modèle est présentée en figure 4.

On peut également exprimer la variable qualitative "mois" selon des fonctions sinus et cosinus et le modèle ci-dessus est équivalent au modèle défini par l'interaction entre le polynôme de degré 5 en t et les variables $\cos(2\pi\ j\ t/T)\ \ j = 1...6$ et $\sin(2\pi\ j\ t/T)\ \ j = 1...5$

Ce résultat est entièrement lié au fait qu'une série de période T connue peut s'exprimer sous la forme d'une combinaison linéaire de fonctions sinus et cosinus dont les périodes sont diviseurs de T

Un inconvénient des régressions sur polynômes est associé aux problèmes d'ajustement en début et fin de série, cet inconvénient est moins présent avec les méthodes d'estimation fonctionnelle. Par contre le modèle paramétrique linéaire présente l'intérêt d'évaluer des hypothèses pouvant être aisément énoncées.

On peut ainsi se poser la question de savoir si tous les mois d'une même saison sont semblables. le modèle qui s'en déduit s'identifie au moyen de contraintes linéaires très simples, "des égalités de moyennes".

Mais on peut aussi tirer parti de la formulation faisant appel aux fonctions sinus et cosinus, en testant l'absence d'harmoniques, ou avec des méthodes de régression pas à pas.

Nous avons ainsi calculé le périodogramme de la série. Il apparaît (figure 5) que deux harmoniques s'imposent correspondant aux périodes 12 et 6 mois. Nous avons donc refait l'ajustement en ne gardant que ces deux harmoniques. Le résultat est présenté sur la figure 6.

Il convient de noter que les ajustements ainsi obtenus par le modèle STL, le modèle linéaire combinant un effet mois et un polynôme de de degré 5 ou le modèle linéaire combinant des fonctions sinus et cosinus de période 12 et 6 avec un polynôme de degré 5 sont quasiment équivalents, les sommes des carrés des écarts entre valeurs observées et estimées étant respectivement égales à 79.5, 71.26 et 82.3, la variance totale étant quant à elle de 777.0. Le carré moyen résiduel associé au modèle simplifié est de plus inférieur au carré moyen du modèle incluant les 11 fonctions sinus et cosinus.

BIBLIOGRAPHIE

CLEVELAND (R.B.), CLEVELAND (W.S.), MCRAE (J.E.) & TERPENNING (I.) 1990 - STL : a seasonal trend decomposition procedure based on Loess. *Journal of Official Statistics.*

PEZENNEC (O.) & BARD (F.X.), 1992 - Importance écologique de la petite saison d'upwelling ivoiro-ghannéenne et changements dans la pêcherie de *Sardinella aurita. Aquat. Living Resourc.* 5 : 249-259.

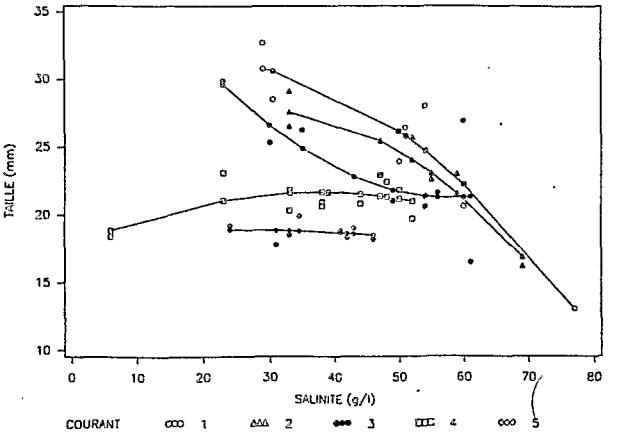

Figure 1 : Valeurs observées et ajustées avec un modèle plynômial de degré 2 selon chaque niveau de courant

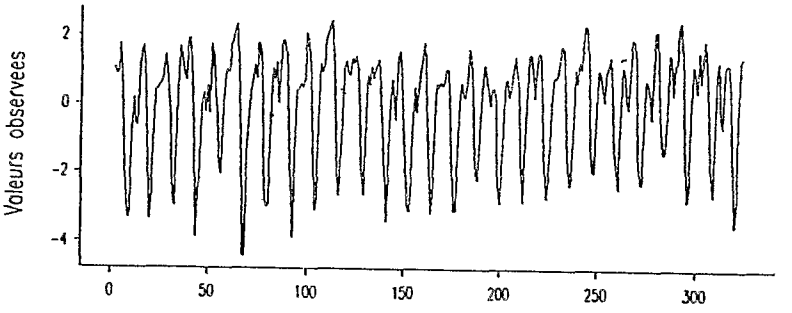

Figure 2 : Données mensuelles de températures de surface, côte ouest de la Côte d'Ivoire (données centrées, en degré celsius)

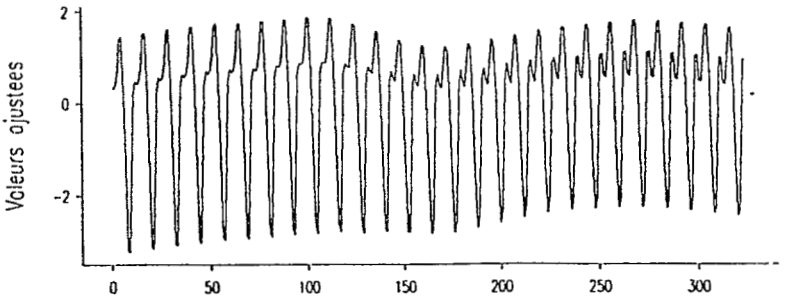

Figure 3 : Valeurs ajustées par la procédure "STL"

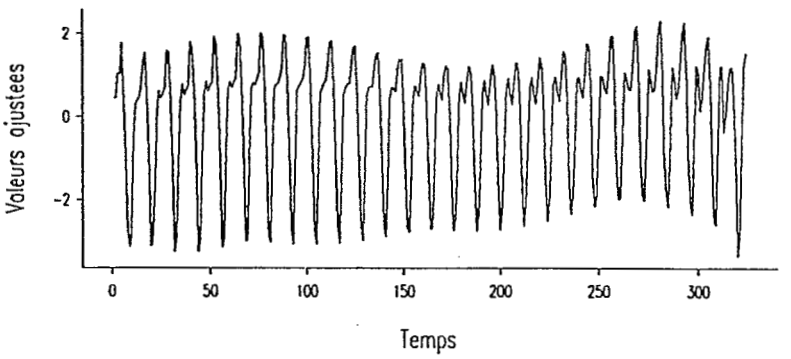

Figure 4 : Valeurs ajustées avec un polynôme de degré 5 pour chacun des douze mois

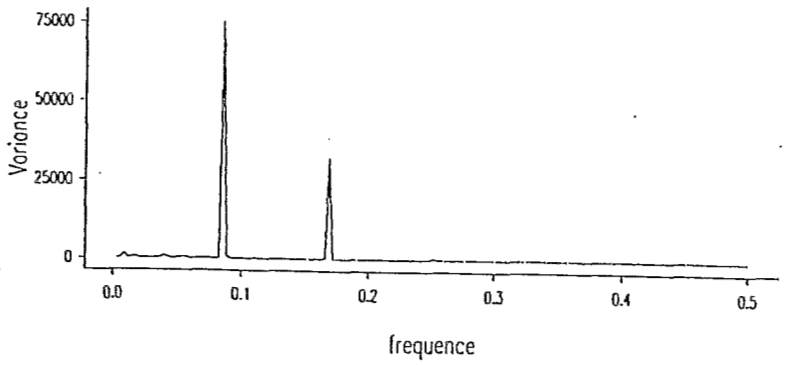

Figure 5 : Périodogrammes (les pics correspondant à des périodes annuelles et semestrielles)

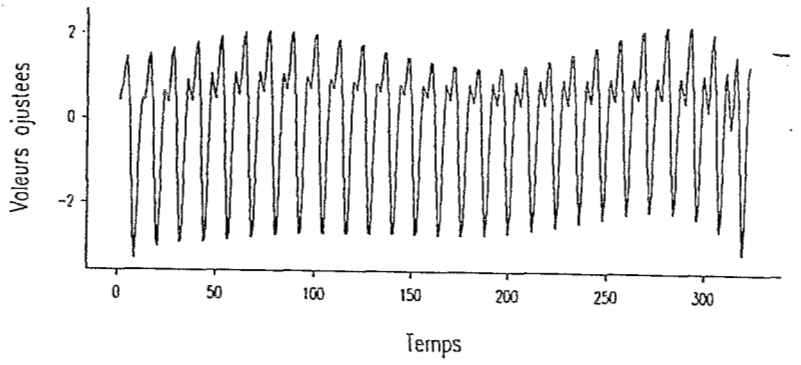

Figure 6 : Valeurs ajustées selon un polynôme de degré 5 en intéraction avec des sinus et cosinus de période 12 et 6 mois