Problème de mauvaise spécification de la variance d'un modèle linéaire

Application à la modélisation de séries temporelles de richesse en sucre de la canne sur l'île de la Réunion

SABINE LAURENT[1], GILLES DUCHARME[2] et PHILIPPE LETOURMY[1]

1 : CIRAD/CA UPR : 13
TA 70/07
Avenue d'Agropolis
34398 Montpellier

2 : Professeur
Institut de mathématiques et de modélisation de Montpellier, cc 051
Université Montpellier II
Place Eugène Bataillon
34095, Montpellier, Cedex 5, France

sabine.laurent@cirad.fr, ducharme@math.univ-montp2.fr et philippe.letourmy@cirad.fr

Résumé **— Ce document présente, dans le cadre d'une étude en mathématiques appliquées sur des données longitudinales de richesse en sucre de la canne, le passage de l'application à la résolution d'un problème théorique original. Nous montrons que la question posée par l'application est souvent en relation avec un domaine théorique large. Un effort de compréhension du contexte de l'étude doit être mené pour permettre la mise en évidence d'une problématique théorique précise à résoudre. L'enjeu d'une thèse appliquée est de permettre d'apporter les solutions appropriées au contexte réel par la résolution d'un problème théorique.**

Mot clés **— Courbe d'évolution, Canne à sucre, Modèle linéaire, Variance**

1. INTRODUCTION ET PROBLÈME POSÉ

La culture de la canne à sucre occupe une place prépondérante dans l'agriculture Réunionnaise en étant la deuxième production de l'île et en participant au maintien d'un grand nombre d'emplois.
Depuis les années 1990, pour rentabiliser la filière, des études agronomiques physiologiques (Martiné, 2003), pédoclimatiques (Gaudin, 1999), des techniques agricoles et organisationnelles sont menées au sein du CIRAD. L'une d'elle concerne l'organisation de l'approvisionnement des bassins canniers réunionnais en étudiant leur fonctionnement et en testant les améliorations possibles (Lejars et al. 2003).
Chaque bassin est caractérisé par un ensemble d'exploitations qui produisent les cannes brutes, des centres de réception qui peuvent s'apparenter à des lieux de stockage, et une usine qui permet la transformation en sucre des cannes. L'organisation de l'approvisionnement régit les flux ou quantités de canne entre l'exploitation et l'usine en passant par le centre de réception, elle est gérée par un calendrier déterminant les périodes de récolte des cannes et les quantités à couper. Notre étude s'inscrit dans un projet d'optimisation de la production de sucre en fonction du calendrier de la récolte. La production de sucre dépend majoritairement de la masse des cannes récoltées et de la teneur en sucre (saccharose) ou richesse.
La teneur en sucre est mesurée sur un échantillon prélevé sur chacune des livraisons de chaque exploitation pendant toute la campagne de récolte, nous obtenons ainsi pour chaque agriculteur et année de récolte une série temporelle de richesse. Les courbes d'évolution de richesse obtenues à partir de ces données montrent une variation significative de la richesse des cannes au cours de la campagne de récolte. La variation concerne une courbe ayant un maximum qui n'est pas atteint au même moment de la campagne selon la zone de production.
Nous proposons une modélisation statistique des séries temporelles de richesse par zone de production pour prévoir la richesse des cannes livrées tout au long de la campagne et pouvoir programmer la coupe des cannes dans les périodes optimales d'un point de vue de la quantité de saccharose contenu dans celles-ci.

2. LES DONNÉES

La courbe de richesse est fonction de facteurs agronomiques et des conditions climatiques supposées prépondérantes. Afin de répondre au mieux à cette question, des données brutes historiques de richesse sur plusieurs campagnes et des données climatiques susceptibles d'expliquer les variations de richesse au cours de la campagne, ont été mises à ma disposition. Les données brutes sont fournies à l'échelle de la livraison de chaque planteur. Un planteur possède plusieurs parcelles qui ne forment pas un ensemble connexe et la surface représentée par l'ensemble de celles-ci varie d'un planteur à l'autre ce qui induit une hétérogénéité de la variance. De plus, nous ne connaissons pas la parcelle d'où provient la livraison de l'exploitant, donc nous ne pouvons pas affecter la richesse à la parcelle de récolte.

Nous avons donc mené un travail de validation des données brutes par le choix des échelles spatiale et temporelle qui allait nous permettre une interprétation agronomique des courbes de prédiction de la richesse en sucre de la canne.

Nous avons choisi l'échelle temporelle hebdomadaire pour pouvoir obtenir un tableau complet et, comme la gestion de l'approvisionnement se fait à la semaine, ceci faciliterait l'utilisation des résultats. Nous avons construit une nouvelle échelle spatiale : la ZÉCÀS (Zones d'Études Cannes À Sucre) avec la collaboration de J. Parriaud (Parriaud J. 2005) pour le georéférencement. Ce sont des zones agricoles, construites à partir des découpages en terroirs et en parcelles (exploitations) de la surface agricole de l'île déjà existants. Elles sont homogènes d'un point de vue des potentiels agricole, pédologique et climatique et contiennent l'ensemble de toutes les parcelles des planteurs y étant présents. Il existe toujours une variabilité au niveau du nombre de livraisons effectuées et des surfaces agricoles représentées entre les ZÉCÀS. Nous avons réglé ainsi le problème de la localisation des cannes, mais il reste une hétérogénéité de la variance et de la covariance induite par la variabilité des surfaces et livraisons.

Un travail d'interpolation des données climatiques aux échelles de la ZÉCÀS et de la semaine, a été mené afin d'inclure ces données comme des variables explicatives de la richesse.

Le jeu de données final contient, pour les six années de récoltes : les richesses hebdomadaires par ZÉCÀS (qui représentent la variable à expliquer Y) et les semaines de coupe, les données climatiques de pluies, températures, rayonnement et évapotranspiration potentielle (qui sont les variables explicatives).

Les prédictions de richesse sont obtenues par modélisation de la variable à expliquer (richesse), en fonction des variables explicatives temporelle et climatiques mises sous forme d'une matrice X. Les données de richesse sont des données répétées dans le temps puisque nous avons plusieurs semaines par année de récolte pour chaque zone de production. Nous avons donc étudié des données longitudinales corrélées dans le temps.

La question posée était : Comment choisir le modèle pour avoir une prédiction de richesse la plus exacte possible ?

3. LA MODÉLISATION

Une étude précédente (Laurent S. 2003) a montré qu'un modèle de type polynomial ($Y = a_0 + a_1 X_1 + a_2 X_2 + a_{12} X_1^2 + ... + e$), dont les paramètres dépendaient de la zone de production et de l'année de récolte et dont la forme et le degré étaient à déterminer, expliquait correctement les évolutions de richesse au cours du temps pour des zones où la courbe était relativement lisse, mais dès que la courbe devenait erratique, la prédiction devenait imprécise.

Nous avons choisi de conserver la structure de modèle polynomial en utilisant un modèle linéaire $Y = X\beta + e$, où $Y = (Y_1,..,Y_n)$ est le vecteur de la variable à expliquer, X (n,p) est la matrice des variables explicatives, supposée de plein rang, $\beta = (\beta_1,\ldots,\beta_p)$ est le vecteur des p paramètres à estimer et $e = (e_1,..,e_n)$ est le vecteur d'erreur supposé de loi normale $N(0, \sigma^2 \Sigma)$. La structure d'espérance du modèle $X\beta$ restait une fonction linéaire et était supposée bien spécifiée. Les paramètres inconnus du modèle étaient β et σ^2 qui seront estimés, et Σ dont les paramètres seront à estimer. La forme de la matrice Σ est choisie *a priori,* donc une erreur sur la spécification de la variance peut alors être commise. Nous avons étudié les conséquences sur les résultats théoriques et les prédictions du modèle dans le cas où la matrice Σ est *a priori* égale à l'identité. Alors β et σ^2 étaient inconnus et ont été estimés par les estimateurs des moindres carrés ordinaires.

Cet estimateur $\tilde{\beta}$ (égal à $(X'X)^{-1}X'Y$) des paramètres β suit une loi normale :

$$\tilde{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1} X'\Sigma X (X'X)^{-1})$$

Il reste donc sans biais, mais n'est plus de variance minimale (Théorème de Gauss Markov).

L'estimateur $\tilde{\sigma}^2$ de la variance σ^2 est biaisé

$$\tilde{\sigma}^2 = \frac{(Y - X\beta)'(Y - X\beta)}{n - p}$$

de biais et variance:

$$biais(\tilde{\sigma}^2) = \frac{\sigma^2 trace(P\Sigma)}{n - p}$$

$$Var(\tilde{\sigma}^2) = \frac{2\sigma^4 trace(P\Sigma)^2}{n - p}$$

où P est le projecteur ($I_n - X(X'X)^{-1}X'$).

Nous avons aussi montré que $\tilde{\sigma}^2$ suit une somme de loi de Khi-2 pondérée (Box, 1954)

$$\tilde{\sigma}^2 \sim \frac{\sum_{i=1}^{n-p} \lambda_i Z_i}{n-p}$$

avec ,

λ_i valeurs propres de la matrice $\sigma^2 P' \Sigma P$,

Z_i suit une loi du khi-2 à un degré de liberté

Ensuite nous avons fait le test de l'hypothèse:

« H0 : $A\beta = 0$ contre H1 : $A\beta \neq 0$ »

Où A(r,p) matrice supposée de plein rang $r \leq p$ et nous avons établi la statistique de test sous H0 utilisée en pratique :

$$\tilde{F} = \frac{\tilde{Q}/r\sigma^2}{\tilde{\sigma}^2/\sigma^2} = \frac{\tilde{Q}}{r\tilde{\sigma}^2}$$

avec, $\tilde{Q} = (A\tilde{\beta})'(A(X'X)^{-1}A)^{-1}A\tilde{\beta}$.

Ainsi nous avons obtenu un rapport de formes quadratiques où $\tilde{Q}$ suit une loi du Khi-2 à r degrés de liberté r et $\tilde{\sigma}^2$ suit une somme de Khi-2 pondérée. Nous n'étions plus dans le cas classique du rapport de deux formes qui suivent des Khi-2, donc nous ne pouvions pas conclure sur la loi de la statistique de test $\tilde{F}$. Le problème théorique que nous avions à résoudre a été d'établir ou d'approximer la loi qui régit ce rapport.

4. CONCLUSION :

Ce document montre la difficulté de mettre en relation, dans certains cas, les questions issues de l'application et les problématiques théoriques. Nous montrons dans notre exemple que le passage de l'application à la théorie demande un investissement particulier des différents acteurs de l'étude. Nous sommes passés d'une question générale concernant la modélisation des courbes de richesse de la canne à sucre à une étude sur l'impact d'une erreur de spécification de la structure de variance d'un modèle linéaire. Nous avons montré que le travail de modélisation statistique se situait au niveau de l'approximation de la loi de la statistique de test pour la sélection des variables explicatives dans le cas d'une mauvaise spécification de la variance.

Nous avons mis en évidence que les clés permettant d'avoir cette relation application/théorie sont que les données soient pertinentes par rapport à la question posée et qu'on puisse les mettre en forme pour faire un travail de recherche statistique dont les résultats répondent à la question posée.

5. RÉFÉRENCES

[1] Box, G E P. (1954) Some theorems on quadratics forms applied in the study of analysis of variance problems. Annals of mathematical statistics, 25, 2, 290-302.

[2] Gaudin and all. (1999) L'eau utile et les caractéristiques hydrodynamiques des sols sous culture de canne à sucre. Agriculture et développement, 30-38.

[3] Laurent, S. (2003) Analyse statistique des courbes de richesse en sucre de la canne pour la gestion de l'approvisionnement d'une usine sur l'île de la Réunion. Rapport de stage de DEA, 48p.

[4] Lejars, C; Letourmy, P; Laurent, S. (2003) Building and assessing supply management scenarios based on cane quality variations : Example of la Reunion Island. South African Sugar Technologists' Association, 580-591.

[5] Martiné, J.F (2003) Modélisation de la production potentielle de la canne à sucre en zone tropicale, sous conditions thermiques et hydriques contrastées. Applications du modèle. Thèse INAPG/Cirad

[6] Parriaud, J. (2005) Les ZÉCÀS : Zones d'Études de la Cannes À Sucre. Rapport d'avancement travaux. Cirad/CA.