Techniques d'enquête, décembre 1996
Vol. 22, n° 2, pp. 177-186
Statistique Canada

Applications du lissage spatial aux données d'enquête

ANN COWLING, RAY CHAMBERS, RAY LINDSAY et BHAMATHY PARAMESWARAN[1]

RÉSUMÉ

Dans le présent article, nous décrivons deux applications du lissage spatial en utilisant des données recueillies dans le cadre d'une enquête économique à grande échelle portant sur les exploitations agricoles australiennes: la première pour les petites régions et la seconde pour les grandes. Dans le premier cas, (petites régions), nous décrivons comment le lissage spatial des poids de l'échantillon peut permettre d'améliorer les estimations. Dans le second cas (grandes régions), nous proposons une méthode de lissage spatial et une méthode de cartographie des données lissées. La méthode standard de pondération utilisée pour l'enquête est une variante de la pondération à régression linéaire. Pour les petites régions, cette méthode est modifiée par l'introduction d'une contrainte sur la variabilité spatiale des poids. Les résultats d'une étude empirique à petite échelle laissent constater que cette méthode réduit comme prévu la variance des estimateurs des petites régions, mais au coût d'une augmentation de leur biais. Pour l'application aux grandes régions, nous décrivons la méthode de régression non paramétrique utilisée pour le lissage spatial des données d'enquête, ainsi que les techniques de cartographie de ces données lissées fondées sur un système d'information géographique (SIG). Nous présentons en outre les résultats d'une étude de simulation réalisée afin de déterminer la méthode et le degré de lissage les plus appropriés pour l'utilisation avec les cartes.

MOTS CLÉS: Estimations pour les petites régions; pondération d'enquête; estimation du noyau; cartographie des données d'enquête.

1. INTRODUCTION

Le Australian Bureau of Agricultural and Resource Economics (ABARE) est une organisation de recherches économiques appliquées rattachée au Department of Primary Industries and Energy. Le ABARE est notamment responsable des enquêtes annuelles conduites auprès d'industries agricoles australiennes choisies qui lui procurent une vaste gamme d'informations sur les caractéristiques économiques et physiques des exploitations agricoles.

La plus importante de ces enquêtes est le Australian Agricultural and Grazing Industries Survey (AAGIS), qui porte sur les exploitations agricoles dont la valeur estimée des opérations (EVAO) s'établit à $22,500 ou plus pour la période visée par le dernier recensement agricole et qui font partie d'une des grandes industries de production agricole (broadacre industries): production céréalière, production de bovins de boucherie et production d'ovins et de laine. Au cours des deux dernières années, environ 1,650 de ses exploitations ont été incluses dans l'échantillon du AAGIS, lequel est stratifié par région géographique, par industrie et par EVAO. Les exploitations échantillonnées sont inégalement réparties sur le territoire australien. Les renseignements normalement recueillis comprennent la latitude et la longitude des exploitations échantillonnées (correspondant à l'emplacement de l'entrée principale). Cette connaissance de l'emplacement des exploitations échantillonnées autorise l'utilisation des méthodes de lissage spatiale décrites plus loin dans le présent article.

Traditionnellement, les estimations du AAGIS ont été présentées uniquement sous forme de tableaux de moyennes calculées pour toute l'Australie, pour chaque État et pour chaque industrie à l'intérieur des États. Toutefois, les préoccupations suscitées au sein de l'industrie rurale et des gouvernements par l'incidence combinée des sécheresses qui ont sévi dans certaines régions de l'Australie et du déclin des prix de certains produits ont mis en lumière l'importance de recueillir des informations plus pertinentes et plus détaillées sur les tendances régionales affichées par le rendement des exploitations. En particulier, on a relevé la nécessité d'obtenir des informations sur la distribution spatiale du rendement des exploitations qui refléteraient la variabilité réelle du climat et de la production d'un bout à l'autre du territoire australien.

La cartographie des variations régionales du rendement économique des exploitations étudiées constitue une forme très efficace de présentation spatiale des informations. Nous utilisons une méthode de régression non paramétrique pour réaliser le lissage spatial des données d'enquête provenant de chacune des exploitations, pour les présenter ensuite sous forme de cartes. Les améliorations récentes apportées à la puissance d'analyse et la disponibilité de systèmes SIG abordables et de haute qualité ont fait de cette forme de présentation une solution de rechange pratique à la méthode traditionnelle de présentation des résultats d'enquête sous forme de tableaux.

Les cartes constituent une méthode pratique de présentation des données pour un certain nombre de raisons. Premièrement, les données présentées sous cette forme s'interprètent rapidement; confronté à un trop grand nombre de tableaux, le client risque de négliger certaines variations locales ou de se sentir «submergé» par les chiffres. Deuxièmement, grâce aux cartes, le client peut plus facilement établir un lien entre

[1] Ann Cowling, CSIRO Division of Fisheries, GPO Box 1538, Hobart TAS 7001, Australie et Australian Bureau of Agricultural and Resource Economics; Ray Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton S017 1BJ, Royaume-Uni; Ray Lindsay et Bhamathy Parameswaran, Australian Bureau of Agricultural and Resource Economics, GPO Box 1563, Canberra ACT 2601, Australie.

la variation géographique d'une variable donnée et celle d'une autre variable. Finalement, les cartes en couleur ont une incidence extrêmement positive sur la présentation visuelle des données.

La demande accrue pour des informations spatiales nous a poussés à porter une attention particulière aux estimations portant sur des petites régions. Pour procéder à de telles estimations (qui découlent naturellement du lissage des données d'enquête aux fins de la présentation dans les cartes) on peut par exemple procéder au lissage spatial des poids d'échantillons. On réduit ainsi la variabilité des estimations portant sur de petites régions.

Dans la section 2, nous présentons une méthode qui permet d'intégrer l'emplacement géographique aux méthodes de pondération des enquêtes du ABARE afin de réduire la variabilité de nos estimations pour les petites régions. Dans la section 3, nous appliquons cette méthode aux estimations sous-régionales correspondant à deux régions agricoles. Dans la section 4, nous décrivons comment les techniques de régression du noyau peuvent servir à produire des cartes qui donneront une bonne indication de la variation géographique locale des variables d'enquête. Nous décrivons deux méthodes de cartographie des données lissées à l'aide du système d'information géographique ARC/INFO. Nous résumons finalement en annexe les résultats d'une étude de simulation comparant diverses méthodes de régression du noyau aux fins de l'utilisation dans les cartes du ABARE.

2. ESTIMATION POUR LES PETITES RÉGIONS AVEC LISSAGE SPATIAL DES POIDS D'ÉCHANTILLON

La méthode normalement utilisée par le ABARE pour le calcul des poids des échantillons est décrite par Bardsley et Chambers (1984). Elle est fondée sur l'hypothèse qui veut qu'à un certain degré d'agrégation (p. ex., région agricole), la variable Y obéira à un modèle linéaire de la forme

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{V} \tag{2.1}$$

où $\boldsymbol{Y}$ est le N-vecteur des valeurs de Y à ce niveau d'agrégation, $\boldsymbol{X}$ est une matrice $N \times p$ des valeurs d'un ensemble de p variables repères, β est un p-vecteur de coefficients de régression inconnu et $\boldsymbol{V}$ est un N-vecteur d'erreurs satisfaisant aux équations $E(\boldsymbol{V}) = \boldsymbol{0}$ et $\text{var}(\boldsymbol{V}) = \sigma^2\Omega$, où σ est un paramètre scalaire inconnu et Ω est une matrice diagonale connue $N \times N$ ayant pour éléments la mesure de l'importance de chaque exploitation, c'est-à-dire la EVAO définie dans la section précédente.

Puisque ce modèle est multivalent, le même groupe de variables repères étant utilisé pour chaque variable de l'enquête, la dimension colonne p de $\boldsymbol{X}$ est habituellement grande. Typiquement, $\boldsymbol{X}$ représentera entre 3 et 7 variables liées aux principaux produits agricoles des exploitations de la région ainsi que des variables factices indiquant la strate industrielle à l'intérieur de la région. La meilleure estimation linéaire non biaisée du total de la population d'une variable d'enquête établie à partir d'un modèle à ce point hyperspécifié donnera généralement des poids extrêmement variables et souvent négatifs.

Comme l'expliquent Bardsley et Chambers (1984), il convient d'éviter les poids négatifs dans une enquête multivalente comme celle du AAGIS. De tels poids peuvent notamment conduire à des estimations négatives de quantités intrinsèquement positives. Ce problème a très souvent été évoqué dans la documentation spécialisée (voir par exemple Deville et Särndal 1992; Bankier, Rathwell et Majkowski 1992; et Fuller, Loughin et Baker 1994). La méthode utilisée par le ABARE pour garantir des poids d'échantillons positifs est fondée sur une modification du type par écrêtage des poids de la meilleure estimation linéaire non biaisée, tel que le suggèrent Bardsley et Chambers (1984).

Étant donné un échantillon de taille n provenant d'une région particulière, la méthode de pondération par écrêtage détermine la valeur du vecteur $\boldsymbol{w}$ du poids de l'échantillon en minimisant le critère de l'erreur quadratique moyenne:

$$Q = \lambda^{-1}\boldsymbol{B}^T\boldsymbol{C}\boldsymbol{B} + (\boldsymbol{w} - \boldsymbol{1})^T\omega(\boldsymbol{w} - \boldsymbol{1}). \tag{2.2}$$

Où $\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{w}$ est un p-vecteur des biais repères correspondant aux différences entre les totaux $\boldsymbol{T}$ de la population (connue) des variables repères p représentées par $\boldsymbol{X}$ et où les estimations d'enquête correspondantes $\boldsymbol{x}^T\boldsymbol{w}$ de ces totaux, représentées par $\boldsymbol{C}$, constituent une matrice diagonale $p \times p$ de «coûts» relatifs non négatifs associés à ces biais, ω est une composante échantillon de Ω, $\boldsymbol{x}$ est une composante échantillon de $\boldsymbol{X}$, $\boldsymbol{1}$ est un n-vecteur de chiffres 1 et λ est une constante scalaire choisie par l'analyste d'enquête. La valeur de $\boldsymbol{w}$ qui minimise Q est

$$\boldsymbol{w} = \boldsymbol{1} + \omega^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\omega^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{1}). \tag{2.3}$$

La constante scalaire λ est le paramètre d'écrêtage associé à ces poids. À mesure que la valeur de λ augmente à partir de zéro, les poids de l'échantillon dans $\boldsymbol{w}$ s'éloignent de leurs meilleures valeurs linéaires non biaisées en vertu du modèle (2.1) (c.-à-d., de leurs valeurs lorsque $\lambda = 0$) et voient leur variabilité diminuer graduellement. Autrement dit, à mesure que la valeur de λ augmente, les variances des estimations de l'enquête fondées sur ces poids diminuent. Cependant, à mesure que la valeur de λ augmente, ces estimations deviennent plus baisées en vertu de (2.1), de sorte que les composantes de $\boldsymbol{B}$ s'éloignent de leurs valeurs zéro à $\lambda = 0$ (où les poids d'échantillon définissent des estimations non biaisées en vertu de (2.1)). Ces composantes voient leur valeur augmenter graduellement (en termes absolus) à mesure que λ augmente.

L'analyste d'enquête cherche une solution de compromis entre ces deux sources concurrentes d'«erreurs» en choisissant la plus petite valeur de λ qui permettra aux poids de l'échantillon dans $\boldsymbol{w}$ de se stabiliser à des valeurs strictement positives le plus proches possible de leurs meilleures valeurs linéaires non biaisées en vertu de (2.1) Ceci permet d'assurer que les composantes de $\boldsymbol{B}$ seront les plus petites possibles, sous réserve de cette exigence de stabilité. Le ABARE choisit cette valeur de λ de manière que les poids des échantillons soient au moins égaux à l'unité.

Les travaux récents du ABARE portant sur l'estimation pour les petites régions ont surtout cherché à modifier cette méthode de pondération par écrêtage afin de créer des poids d'échantillons dont la variabilité spatiale serait moindre. Nous y parvenons en modifiant le critère Q de l'erreur quadratique moyenne dans (2.2) afin d'inclure une contrainte de la variabilité spatiale, tout en continuant de considérer les éléments de la variable $\boldsymbol{Y}$ comme des variables indépendantes.

Désignons par $\boldsymbol{K}$ une matrice $n \times n$ reflétant la distance Euclidienne entre les exploitations agricoles échantillonnées, de sorte que $\boldsymbol{K}$ soit symétrique et non négative, $\boldsymbol{K}_{ii} = 1$ pour toutes les valeurs de i et $\boldsymbol{K}_{ij} \downarrow 0$ à mesure que la distance qui sépare l'exploitation i de l'exploitation j augmente. Supposons que $\boldsymbol{u} = \boldsymbol{w} - \mathbf{1}$. La valeur de $\boldsymbol{u}$ que nous choisirons devra faire en sorte que lorsque $\boldsymbol{K}_{ij}$ est grand, la différence entre $\boldsymbol{u}_i$ et $\boldsymbol{u}_j$ est petite. Autrement dit, nous chercherons à minimiser une quantité de la forme

$$\sum_{i \in s} \sum_{j \in s} \boldsymbol{K}_{ij}(\boldsymbol{u}_i - \boldsymbol{u}_j)^2 = 2(\boldsymbol{u}^{(2)})^T \boldsymbol{K}\mathbf{1} - 2\boldsymbol{u}^T \boldsymbol{K}\boldsymbol{u} \qquad (2.4)$$

où $(\boldsymbol{u}^{(2)})_i = (\boldsymbol{u}_i)^2$. Une modification appropriée du critère de l'erreur quadratique moyenne (2.2) conduit à la minimisation de

$$Q^* = \lambda^{-1}\boldsymbol{B}^T\boldsymbol{C}\boldsymbol{B} + \boldsymbol{u}^T\omega\boldsymbol{u} + (\boldsymbol{u}^{(2)})^T\boldsymbol{K}\mathbf{1} - \boldsymbol{u}^T\boldsymbol{K}\boldsymbol{u}.$$

La minimisation par rapport à $\boldsymbol{u}$ conduit à

$$\boldsymbol{u} = \eta^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\eta^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\mathbf{1})$$

à condition qu'il existe une valeur η^{-1}, où

$$\eta = \text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K} + \omega. \qquad (2.5)$$

Il apparaît ainsi clairement que

$$\boldsymbol{w} = \mathbf{1} + \eta^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\eta^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\mathbf{1}). \qquad (2.6)$$

On constate que le critère modifié de l'erreur quadratique moyenne Q^* pondère à la fois le critère de lissage spatial donné dans (2.4) et le terme correspondant à la variance de l'erreur de prédiction des estimations de l'échantillon, $\boldsymbol{u}^T\omega\boldsymbol{u}$. Comme l'échelle de $\boldsymbol{K}$ a été spécifiée arbitrairement, la pondération comparative des deux critères doit être modifiée en multipliant la matrice spatiale $\{\text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K}\}$ par un facteur ϕ, afin de la rendre comparable par la taille à la matrice d'hétéroscédasticité ω, et en ajoutant un paramètre α, $0 \le \alpha \le 1$, à l'expression de η dans l'équation (2.5) de sorte que

$$\eta = (1 - \alpha)\phi\{\text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K}\} + \alpha\omega.$$

On peut dériver ces poids d'échantillons à lissage spatial selon une autre méthode qui donne un meilleur aperçu de la façon dont ils devraient être interprétés. Cette méthode découle du fait que la matrice

$$\eta = \begin{bmatrix} \sigma_1^2 + \sum_{m \neq 1} \boldsymbol{K}_{1m} & -\boldsymbol{K}_{12} & \cdots & -\boldsymbol{K}_{1n} \\ -\boldsymbol{K}_{21} & \sigma_2^2 + \sum_{m \neq 2} \boldsymbol{K}_{2m} & \cdots & -\boldsymbol{K}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{K}_{n1} & -\boldsymbol{K}_{n2} & \cdots & \sigma_n^2 + \sum_{m \neq n} \boldsymbol{K}_{nm} \end{bmatrix}$$

peut être assimilée à l'équation $\eta = \boldsymbol{S}\boldsymbol{R}\boldsymbol{S}$, où $\boldsymbol{S}$ est une matrice diagonale avec $\boldsymbol{S}_{ii} = (\sigma_i^2 + \sum_{m \neq i} \boldsymbol{K}_{im})^{1/2}$, et $\boldsymbol{R}$ est une matrice de corrélation avec

$$R_{ij} = \begin{cases} 1 & \text{si } i = j \\ -K_{ij}\left\{\left(\sigma_i^2 + \sum_{m \neq i} K_{im}\right)\left(\sigma_j^2 + \sum_{m \neq j} K_{jm}\right)\right\}^{-1/2} & \text{si } i \neq j. \end{cases}$$

Ainsi, les poids d'échantillons à lissage spatial peuvent être dérivés comme des poids de régression de type à écrêtage en présumant que la variable $\boldsymbol{Y}$ obéit à un modèle linéaire de la forme (2.1), avec une valeur de $\boldsymbol{V}$ redéfinie pour satisfaire $E(\boldsymbol{V}) = \mathbf{0}, \text{var}(\boldsymbol{Y}_i) = \sigma_i^2 + \sum_{m \neq i} \boldsymbol{K}_{im}$, et $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = -\boldsymbol{K}_{ij}$ pour $i \neq j$. La méthode habituelle de pondération par écrêtage mène ainsi directement à (2.6), avec une valeur de η définie par (2.5). À noter qu'en vertu de ce modèle, on obtient une corrélation négative entre les exploitations agricoles voisines.

Cette deuxième méthode de dérivation montre clairement que le recours au lissage spatial pour les poids de l'enquête s'accorde mal avec les notions standards d'efficacité statistique, dans la mesure où on s'intéresse à l'estimation au niveau de l'agrégat. Comme la corrélation spatiale entre les exploitations voisines sera typiquement positive, l'estimation efficace au niveau de l'agrégat comportera nécessairement une pondération fondée sur (2.3) où ω sera remplacé par une matrice non diagonale variance/covariance reflétant cette corrélation spatiale positive. Ces poids ne sont pas ceux que l'on obtient lorsqu'on impose une contrainte de similarité spatiale. En conséquence, il faudra s'attendre à ce que ces poids «efficaces pour les grandes régions» tendent à être plus différents pour les exploitations voisines qu'ils ne le seront pour les exploitations éloignées l'une de l'autre. Autrement dit, il y a un prix à payer pour la pondération: si nous avons besoin d'estimations au niveau de l'agrégat moins variables, ceci aura tendance à conduire à des estimations plus variables pour les petites régions. Inversement, si on choisit l'équation (2.6) aux fins de la pondération à cause de ses propriétés souhaitables pour les petites régions, il faudra s'attendre à ce que les estimations au niveau de l'agrégat obtenues en faisant la somme de ces estimations pour les petites régions soient moins efficaces.

Les poids d'échantillons à lissage spatial (2.6) ont été appliqués à l'aide de l'équation

$$\boldsymbol{K}_{ij} = \exp(-d\|\boldsymbol{z}_i - \boldsymbol{z}_j\|), \qquad (2.7)$$

où $\|z_i - z_j\|$ représente la distance qui sépare l'exploitation i de l'exploitation j, et d est une constante de contrôle du rayon du cercle entourant la i-ième exploitation, à l'intérieur duquel on a procédé au lissage spatial. Plus la valeur de d est petite, plus grand sera le rayon du lissage spatial. À l'heure actuelle, la constante de multiplication ϕ correspond au rapport des déterminants des matrices $\boldsymbol{K}$ et ω, élevées à la puissance n^{-2}. Nous procédons, à la section suivante, à une évaluation empirique de cette méthode.

3. APPLICATION DE LA PONDÉRATION D'ÉCHANTILLON À LISSAGE SPATIAL

Les résultats initiaux d'une évaluation de la première méthode de pondération par écrêtage à lissage spatial décrits dans la section précédente sont présentés aux tableaux 1 à 3. Ils correspondent à deux régions agricoles distinctes. La première, appelée A, se trouve en Nouvelle-Galles du Sud. En termes d'espace, cette région est relativement homogène, étant située dans la portion sud-ouest de l'État. La production de blé et de riz et la production de laine et d'agneau y sont les principales activités agricoles. La deuxième région, appelée B, se trouve en Australie-Occidentale. C'est une région plus hétérogène, comportant des exploitations de productions végétales et de production de laine, dans le centre-ouest de l'État, et des exploitations beaucoup plus vastes d'élevage du bétail et de cultures installées sur des terres agricoles marginales dans le sud-est. On y produit principalement du blé et des légumineuses, ainsi que de la laine.

Nous avons utilisé six variations des poids d'écrêtage à lissage spatial (2.6) avec une valeur de $\boldsymbol{K}$ donnée par (2.7), définie par les valeurs de $d = 0.05$ (effets spatiaux faibles) et $d = 0.005$ (effets spatiaux forts), et avec des valeurs de $\alpha = 0.9$ (accent sur les poids d'écrêtage standards), $\alpha = 0.5$ (importance égale accordée aux poids d'écrêtage standards et aux poids à lissage spatial) et $\alpha = 0.1$ (accent sur les poids à lissage spatial).

Le tableau 1 présente les valeurs relatives du biais associées à l'estimation des valeurs repères totales liées au produit principal pour chaque région, en vertu de ces différents systèmes de pondération, ainsi que les valeurs correspondantes du biais associées aux poids d'écrêtage standards. L'augmentation de la valeur de ces biais à mesure qu'on augmente le lissage spatial des poids est évidente. Comme il existe une corrélation positive entre ces repères de production et la plupart des variables économiques mesurées dans le cadre de l'enquête, on peut s'attendre à ce que ces biais repères se traduisent par une augmentation correspondante des biais des estimations de l'enquête fondés sur ces poids.

Les figures 1 à 4 illustrent les différences qui existent entre les poids lissés et les poids d'écrêtage standards pour les deux combinaisons «extrêmes» de α et d dans les deux régions, qui changent à mesure que change la taille des exploitations échantillonnées (mesure correspondant au logarithme de la valeur estimée des opérations agricoles, ou log(EVAO)).

Soulignons qu'un lissage spatial relativement fort (figures 1 et 3) a pour effet d'accroître les poids de la plupart des exploitations les plus grandes de l'échantillon, tout en diminuant radicalement les poids d'un petit nombre d'exploitations plus petites. Le lissage spatial faible (figures 2 et 4) influe beaucoup moins sur les poids, et le rapport entre la taille de l'exploitation et la direction du changement des poids n'est pas très évident. En conséquence, on devrait s'attendre

Tableau 1

Valeurs (en pourcentages relatifs) des biais associés à l'estimation des variables repères correspondant aux principaux produits agricoles de la région A (taille de l'échantillon $n = 101$ exploitations) et de la région B (taille de l'échantillon $n = 85$ exploitations) obtenues avec des poids d'écrêtage standards (2.3) et des poids d'écrêtage à lissage spatial (2.6)

		Blé	Moutons	Riz
Région A				
Poids d'écrêtage standards		-0.50	5.0	13.0
Poids d'écrêtage à lissage spatial				
$d = 0.05$	$\alpha = 0.9$	-0.50	4.6	11.9
	$\alpha = 0.5$	-0.46	4.7	12.4
	$\alpha = 0.1$	0.07	6.2	17.4
$d = 0.005$	$\alpha = 0.9$	-0.40	4.9	12.7
	$\alpha = 0.5$	0.80	8.9	28.0
	$\alpha = 0.1$	9.20	25.0	60.0
		Blé	Moutons	Légumineuses
Région B				
Poids d'écrêtage standards		0.43	-1.25	1.49
Poids d'écrêtage à lissage spatial				
$d = 0.05$	$\alpha = 0.9$	0.42	-1.16	1.37
	$\alpha = 0.5$	0.44	-1.14	1.40
	$\alpha = 0.1$	0.69	-1.25	2.53
$d = 0.005$	$\alpha = 0.9$	0.50	-1.20	1.68
	$\alpha = 0.5$	1.51	1.14	9.73
	$\alpha = 0.1$	26.57	19.61	45.46

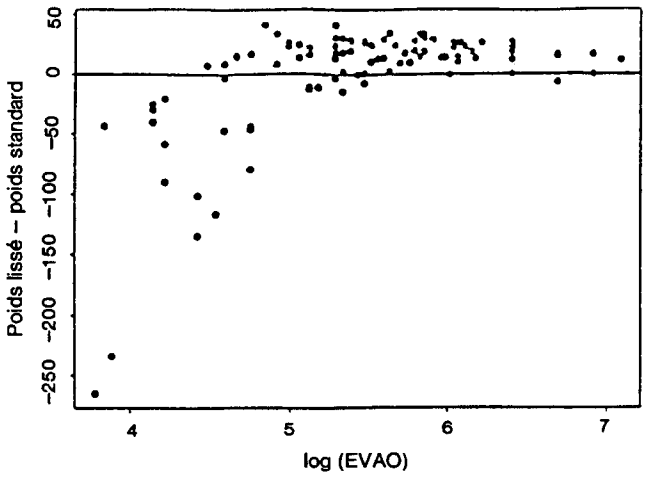

Figure 1. Différence entre le poids lissé avec $\alpha = 0.1$ et $d = 0.005$ et le poids d'écrêtage standard, Région A

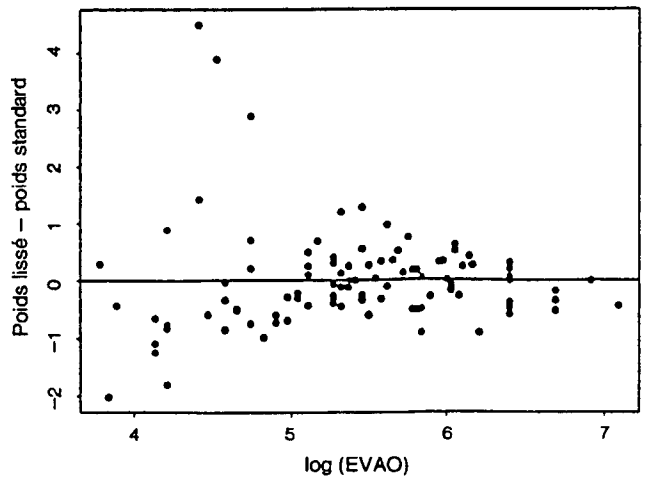

Figure 2. Différence entre le poids lissé avec $\alpha = 0.9$ et $d = 0.05$ et le poids d'écrêtage standard, Région A

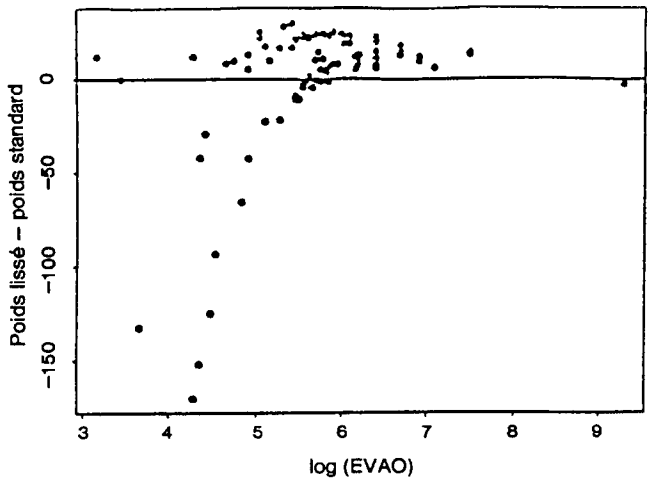

Figure 3. Différence entre le poids lissé avec $\alpha = 0.1$ et $d = 0.005$ et le poids d'écrêtage standard, Région B

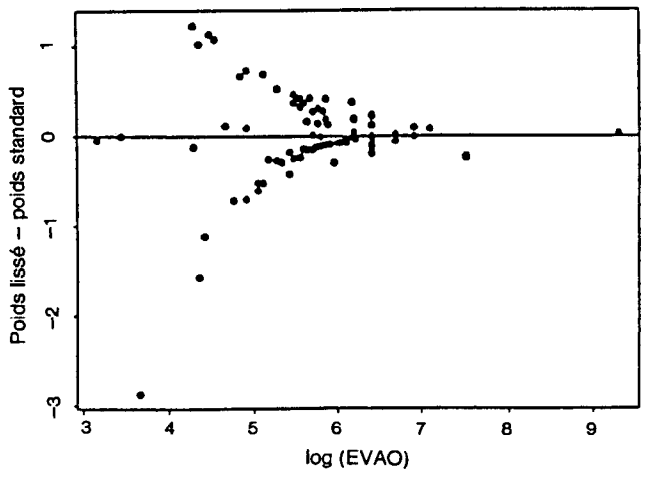

Figure 4. Différence entre le poids lissé avec $\alpha = 0.9$ et $d = 0.05$ et le poids d'écrêtage standard, Région B

à un changement à la hausse des estimations des enquêtes pour ces régions avec l'introduction de poids d'échantillons à fort lissage spatial. Compte tenu du biais positif accru indiqué au tableau 1, cette augmentation devrait être essentiellement due à l'introduction d'un biais positif dans ces estimations.

Cette augmentation du biais est-elle compensée par un écart-type plus faible? Pour répondre à cette question, nous avons calculé les valeurs estimées de l'enquête et les écarts-types correspondant à une variable financière clé: le total des coûts au comptant. Ces estimations sont compilées aux tableaux 2 (région A) et 3 (région B). On fournit les estimations pour chacune des régions ainsi que pour les sous-régions comprises à l'intérieur de chacune d'elles et qui sont identifiées dans les tableaux par SR-i, l'indice i variant de 1 à 6 pour la région A et de 1 à 7 pour la région B.

On observe qu'en général, la réponse à la question posée ci-dessus est affirmative. Les valeurs estimées de l'écart-type des estimations de l'enquête diminuent à mesure qu'augmente le degré de lissage spatial des poids (de gauche à droite, dans les tableaux). Toutefois, comme prévu, les estimations augmentent elles aussi en taille, montrant un biais positif de plus en plus important. Dans l'ensemble, l'avantage procuré par une baisse de l'écart-type semble compenser pour l'augmentation du biais, sauf dans les cas de lissage spatial plus important ($\alpha = 0.1$, $d = 0.005$). Dans ce dernier cas, l'augmentation du biais dépasse la réduction de l'écart-type. Le choix d'un $\alpha = 0.1$ et d'un $d = 0.05$ paraît être un compromis acceptable, permettant d'obtenir un équilibre raisonnable (mais non spectaculaire) entre le biais et la variance dans la région A, et peu de changements dans les estimations de la région B.

Tableau 2

Estimations (valeurs estimées de l'écart-type entre parenthèses) de la valeur moyenne de Y = coûts totaux au comptant dans les sous-régions SR-1 à SR-6, constituant la Région A (taille de l'échantillon $n = 101$ exploitations), obtenues avec des poids d'écrêtage standards (2.3) et des poids d'écrêtage à lissage spatial (2.6)

	Poids standards	Poids d'écrêtage à lissage spatial					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR-1	100,618	100,453	101,297	107,263	102,059	112,635	135,419
	(24,551)	(24,511)	(23,906)	(20,487)	(23,474)	(18,923)	(18,011)
SR-2	115,320	115,417	116,002	120,362	116,917	126,165	153,707
	(26,754)	(26,661)	(26,448)	(25,637)	(26,423)	(25,990)	(27,975)
SR-3	167,524	167,453	167,486	168,257	167,709	170,781	187,683
	(28,479)	(28,467)	(28,473)	(28,426)	(28,175)	(26,471)	(24,211)
SR-4	182,940	180,317	177,838	163,556	176,257	174,077	192,296
	(106,471)	(105,485)	(101,012)	(74,418)	(97,823)	(69,109)	(43,651)
SR-5	132,050	132,083	132,389	134,786	132,490	136,369	151,046
	(25,089)	(25,096)	(25,154)	(25,475)	(25,173)	(24,410)	(23,110)
SR-6	132,493	132,184	133,204	141,623	133,763	147,652	192,781
	(44,385)	(44,546)	(44,757)	(46,736)	(45,078)	(46,953)	(53,105)
Région A	134,114	133,807	134,141	137,080	134,506	142,040	166,432
	(15,691)	(15,655)	(15,426)	(13,845)	(15,199)	(13,494)	(12,815)

Tableau 3
Estimations (valeurs estimées de l'écart-type entre parenthèses) de la valeur moyenne de Y = coûts totaux au comptant dans les sous-régions SR-1 à SR-7, constituant la Région B (taille de l'échantillon n = 85 exploitations), obtenues avec des poids d'écrêtage standards (2.3) et des poids d'écrêtage à lissage spatial (2.6)

	Poids standards	Poids d'écrêtage à lissage spatial					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR-1	183,194	183,262	183,528	186,151	184,287	195,138	257,652
	(64,851)	(64,325)	(64,051)	(64,967)	(64,132)	(69,859)	(59,518)
SR-2	261,952	261,487	261,119	261,182	261,938	276,912	331,805
	(70,989)	(70,601)	(70,502)	(73,131)	(70,723)	(79,751)	(67,356)
SR-3	113,499	113,441	113,742	116,847	114,631	125,525	157,007
	(30,304)	(30,289)	(30,255)	(30,731)	(30,377)	(31,507)	(32,500)
SR-4	242,220	242,182	242,208	242,221	242,163	242,439	250,871
	(26,160)	(25,671)	(26,159)	(26,160)	(26,154)	(24,244)	(24,836)
SR-5	134,524	134,970	135,700	139,122	134,734	131,448	148,629
	(32,420)	(32,528)	(32,432)	(30,607)	(32,202)	(27,867)	(27,942)
SR-6	176,540	176,977	175,708	163,241	172,076	148,434	171,856
	(60,377)	(60,703)	(59,214)	(46,361)	(55,925)	(36,218)	(39,527)
SR-7	205,287	205,644	205,433	202,039	204,519	194,998	219,959
	(44,137)	(44,008)	(43,963)	(44,044)	(43,972)	(45,434)	(51,690)
Région B	176,283	176,342	176,397	176,822	176,294	179,998	216,445
	(19,039)	(18,869)	(18,874)	(18,213)	(18,511)	(18,540)	(17,099)

4. ESTIMATION ET CARTOGRAPHIE DES MOYENNES LOCALES

Une carte de données d'enquête est une surface bidimensionnelle qui permet d'estimer la fonction moyenne spatiale de la variable d'enquête dans la population. En pratique, on obtient une telle carte en appliquant des techniques de régression non paramétriques aux données d'enregistrement unitaires pondérées obtenues dans le cadre de l'enquête.

Le ABARE utilise la régression de noyau (une technique non paramétrique) pour produire des cartes qui montrent la variation spatiale des surfaces de la fonction moyenne spatiale estimée de variables d'enquête clés. Ces surfaces sont obtenues en remplaçant les valeurs d'échantillons observées de ces variables par des moyennes localement pondérées. En outre, pour chaque carte de moyenne locale, on produit une carte correspondante qui donne une estimation de la variabilité locale de la variable d'intérêt. Nous présentons ci-après un bref aperçu de la technique: pour des raisons de simplification, nous traiterons uniquement du cas univarié. Voir Ruppert et Wand (1994), Wand et Jones (1995, p.140) et les articles auxquels ils font référence pour en savoir plus sur le cas multivarié.

Nous présumons que la population finie générée est un échantillon iid $\{(Z_i, Y_i), i = 1, ..., N\}$ tiré d'une superpopulation où Y_i est la valeur d'une variable de réponse Y observée au lieu Z_i. Nous supposons que les observations obéissent au modèle

$$Y_i = m(Z_i) + \epsilon_i, \qquad i = 1, ..., N$$

où $m(z) = E(Y \mid Z = z)$ correspond à la moyenne conditionnelle de Y, étant donné Z, et ϵ_i désigne les variables aléatoires indépendantes ayant une moyenne zéro et une variance $\sigma^2(z)$. Supposons que les termes d'erreurs ϵ_i sont indépendants du processus de sélection de l'échantillon, de sorte que les valeurs de l'échantillon $\{(Z_i, Y_i), i = 1, ..., n\}$ obéissent au même modèle, et désignons par f la densité de $Z_1, ..., Z_n$.

Le choix naturel de la moyenne locale d'un point z quelconque devient alors la moyenne des valeurs de la variable réponse pour les observations dont l'emplacement est proche de z, puisque les observations de points éloignés ont tendance à montrer des valeurs moyennes très différentes. La moyenne locale est donc définie comme une moyenne pondérée

$$\hat{m}(z) = n^{-1} \sum_{i=1}^{n} W_i(z)\, Y_i$$

où les poids $\{W_i(z)\}$ dépendent des emplacements $\{Z_i\}$ des observations de l'échantillon, et où $\hat{m}(z)$ est une estimation de $m(z)$.

Les poids sont élaborés à l'aide d'une fonction K connue sous le nom de noyau, continue, bornée, symétrique et dont l'intégrale est 1. Diverses séquences de poids ont déjà été proposées: les poids traditionnels de Nadaraya-Watson (Nadaraya 1964 et Watson 1964) sont

$$W_i(z) = h^{-1} K\{(z - Z_i)/h\} \Big/ \left[(nh)^{-1} \sum_{j=1}^{n} K\{(z - Z_j)/h\} \right],$$

où h est un facteur scalaire appelé largeur de bande. La fonction K donne à une observation proche de z une influence relativement plus grande sur la moyenne locale à cet emplacement qu'elle ne le fait pour une observation plus éloignée de z.

Lorsque le nombre d'observations est limité, une fenêtre à largeur de bande fixe pourrait contenir un nombre limité de points et l'estimateur correspondant risquera donc d'avoir une variance très élevée. On peut éviter ce problème en utilisant la méthode du k-ième plus proche voisin dans laquelle une largeur de bande différente est utilisée pour chaque estimation du point z. La largeur de bande à z est la distance au k-ième plus proche voisin de z, de sorte qu'il y aura toujours exactement k points dans la fenêtre. Désignons par h_k la distance qui sépare z de son k-ième plus proche voisin. Les poids k-ième plus proche voisin de Nadaraya-Watson sont

$$W_{ih_k}(z) = h_k^{-1} K\{(z - Z_i)/h_k\} \Big/ \left[(nh_k)^{-1} \sum_{j=1}^{n} K\{(z - Z_j)/h_k\} \right].$$

Nous présentons au tableau 4 les propriétés de l'erreur quadratique moyenne (EQM) asymptotique des estimateurs habituels (largeur de bande fixe) et des estimateurs k-ième plus proche voisin, tels qu'établies par Härdle (1990, p. 46).

Tableau 4
Biais et variance asymptotiques des estimateurs de Nadaraya-Watson;
$c_K = \int K^2(u)du$, $d_K = \int u^2K(u)du$

	Largeur de bande fixe	k-ième plus proche voisin
Biais	$h^2\frac{(m''f + 2m'f')(x)}{2f(x)}d_K$	$\left(\frac{k}{n}\right)^2\frac{(m''f + 2m'f')(x)}{8f^3(x)}d_K$
Variance	$\frac{\sigma^2(x)}{nhf(x)}c_K$	$\frac{2\sigma^2(x)}{k}c_K$

Il apparaît clairement que le biais de la fonction de régression estimée peut être réduit en utilisant une largeur de bande plus étroite h (nombre de plus proches voisins k), mais ceci conduit à une estimation bruitée $\hat{m}$ dont les détails locaux masquent les caractéristiques globales de la courbe ($\hat{m}$ est assorti d'une grande variance). Si $h(k)$ est grand, $\hat{m}$ sera plus lisse, mais les caractéristiques globales seront amorties ($\hat{m}$ est assorti d'un biais élevé et d'une faible variance). Ainsi, le biais ne peut être réduit qu'aux dépens de la variance, et vice versa, la valeur h de la largeur de bande déterminant le rapport du biais (au carré) sur la variance.

En réalité, le plan d'échantillonnage et la distribution spatiale d'une variable d'enquête Y ne seront pas indépendants, et les moyennes locales simples pour Y dérivées des données de l'échantillon ne permettront pas d'obtenir de bonnes estimations des moyennes de la population locale de cette variable. Pour contourner ce problème, les poids du noyau sont multipliés par les poids de l'enquête pour donner les poids du lissage final utilisés dans le calcul de la moyenne locale. Cette opération équivaut à calculer la valeur estimée de la moyenne de la population locale $m(z)$ de Y, en tenant pour acquis qu'elle est localement linéaire pour les mêmes variables repères que celles utilisées dans la modélisation de la moyenne globale de la population de Y.

Il a déjà été question, dans la documentation spécialisée, d'une vaste gamme de méthodes de lissage du noyau. Outre les diverses séquences de poids de lissage $\{W_i\}$, il existe différents types de largeurs de bandes et de nombreuses méthodes de sélection automatique de la largeur de bande. Nous avons donc procédé à une simulation destinée à déterminer la méthodologie fondée sur le noyau la plus appropriée pour l'établissement des cartes du ABARE. Cette étude est décrite en annexe.

L'incertitude qui entoure l'estimation de la moyenne spatiale dérivée par un lissage spatial fondé sur le noyau peut être illustrée par la cartographie de la variabilité locale de la variable d'intérêt. Les zones de variabilité locale élevée correspondent aux régions où la carte de la fonction moyenne est moins précise, et vice versa pour les zones de variabilité locale faible.

La méthode habituelle de détermination des régions de confiance pour une estimation de la courbe du noyau a reçu le nom de «méthode bootstrap» (voir Härdle 1990; Hall 1992 et les ouvrages cités en référence dans ces articles). Toutefois, pour des raisons d'efficacité de calcul, nous utilisons les «expectiles» (Newey et Powell 1987) de la distribution spatiale de Y pour décrire cette variabilité locale. Un expectile présente un rapport à la moyenne équivalant au rapport du quartile à la médiane. En particulier, la différence entre le 75e et le 25e expectiles d'une distribution constitue une mesure de l'étalement de la distribution comparable à l'interquartile. Le programme de lissage contient un module pour la régression M-quantile non paramétrique (Breckling et Chambers 1988) qu'on utilise pour ajuster la surface lissée aux expectiles de la distribution Y à un endroit donné. La différence entre les surfaces lissées du 75e et du 25e expectiles (l'expectile lissé analogue à l'étendue de l'interquartile) est alors cartographiée pour montrer les zones de haute et de basse variabilité des données.

On constate sans surprise que cette étendue lissée de l'interexpectile tend à être plus élevée dans les régions où les exploitations sont éloignées les unes des autres et où la variabilité de Y est donc la plus élevée. Nous présentons à la figure 6 la carte de l'étendue de l'interexpectile correspondant à la figure 5. Notez que ces cartes de l'étendue de l'interexpectile lissé fournissent des informations semblables aux bandes de confiance à n'importe quel point particulier de la carte.

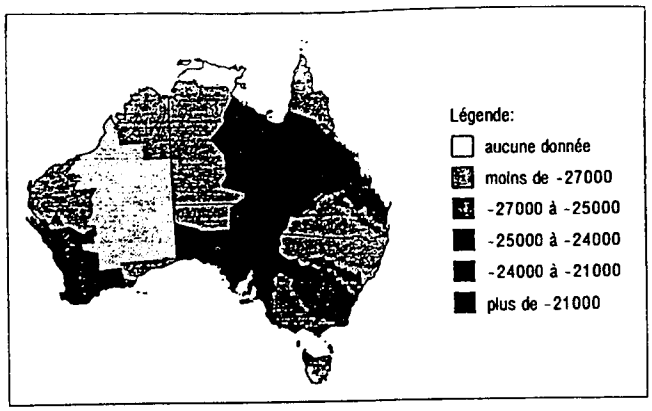

Figure 5. Carte en polygones du profit des exploitations agricoles en 1991-1992, sur l'ensemble des grandes exploitations ($)

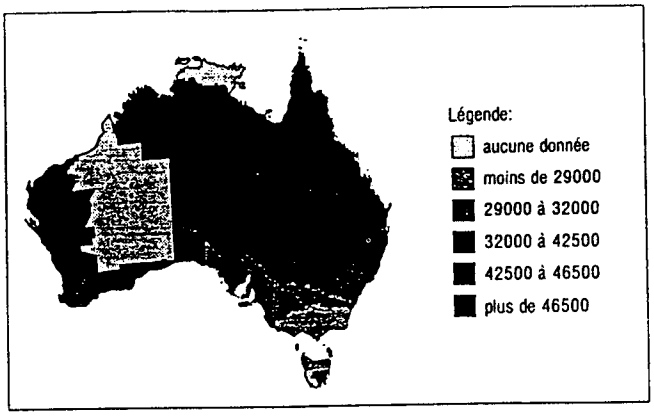

Figure 6. Carte en polygones de l'étendue de l'interexpectile du profit des exploitations agricoles en 1991-1992, sur l'ensemble des grandes exploitations ($)

Toutefois, elles n'offrent pas la même interprétation de l'échantillonnage répété que les intervalles de confiance et devraient donc servir d'indicateurs plutôt que de mesures de l'incertitude associées à une carte particulière.

Pour des raisons de confidentialité, il convient de prendre garde, lors de la préparation des cartes de données lissées aux fins de publication, de ne pas révéler l'emplacement des exploitations étudiées. Il convient en outre d'assurer une qualité des résultats qui soit compatible avec les systèmes d'éditique. Nous avons mis au point deux méthodes permettant de générer les cartes finales répondant à ces exigences en utilisant le SIG ARC/INFO.

Dans la première méthode, un polygone de Thiessen est construit autour de chaque exploitation. Ce polygone délimite l'espace qui se trouve plus proche de la ferme en question que de toutes les autres. La ferme n'est pas au centre de ce polygone, et le périmètre du polygone ne suit pas les limites de l'exploitation, ce qui permet de dissimuler l'emplacement des fermes étudiées comme l'illustre la figure 7. Chaque polygone est coloré selon la valeur lissée de Y à l'emplacement de la ferme dans ce polygone. On utilise habituellement dix couleurs pour chaque carte, et les déciles estimés de la population des données lissées servent de limites à la zone colorée. Les cartes présentées dans le présent article sont des analogues en noir et blanc de ces cartes en couleur.

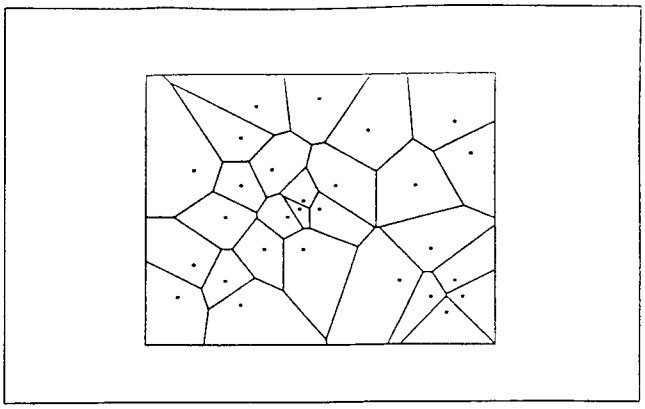

Figure 7. Polygones de Thiessen construits autour d'exploitations choisies pour l'enquête du ABARE. L'emplacement de la ferme est indiqué par un petit carré à l'intérieur de chacun des polygones

Dans la deuxième méthode, les valeurs lissées d'une grille rectangulaire dense remplacent les valeurs lissées à l'emplacement de la ferme et on procède à une interpolation mineure supplémentaire des données dans le système ARC/INFO. On établit une surface continue tridimensionnelle passant par les valeurs lissées des points de la grille, en deux étapes. Dans la première approximation, on construit une surface en facettes triangulaires à l'aide de la triangulation de Delauney pour ajuster ensuite un polynôme bivarié du cinquième degré à l'intérieur de chacun de ces triangles à l'aide de l'algorithme de Akima (Akima 1978). On détermine ensuite les contours de la surface continue résultante en utilisant les déciles estimés de la population. Cette méthode est illustrée à la figure 8.

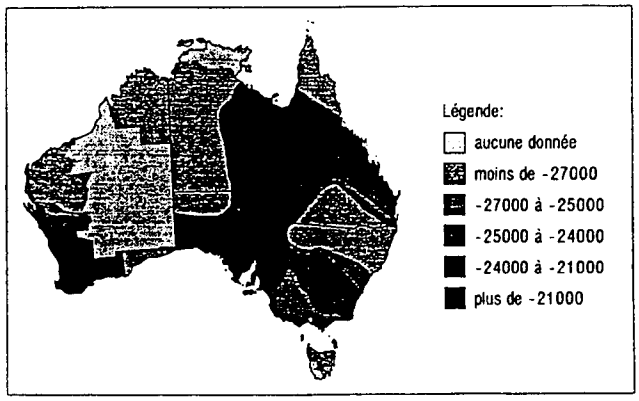

Figure 8. Carte d'isolignes des profits des exploitations agricoles en 1991-1992, pour l'ensemble des grandes exploitations ($)

Avec cette deuxième méthode de présentation, les emplacements des fermes visées par l'enquête ne sont pas pris en compte, ce qui permet de cacher complètement la provenance des données. On obtient en outre des contours plus lisses et les résultats ne sont pas aussi morcelés que sur les cartes en polygones. En outre, le personnel graphique du ABARE préfère cette méthode puisqu'elle réduit le nombre des sections à colorer séparément et qu'elle présente des exigences de stockage moindres, ce qui facilite la manipulation des cartes avec les logiciels d'éditique. Elle a cependant pour inconvénient d'allonger le temps de traitement à l'étape du SIG.

Puisque les méthodes décrites plus haut font l'interpolation des données sur l'ensemble du territoire australien, y compris les zones où on ne pratique pas l'agriculture, l'étape finale de la production des cartes avec le système ARC/INFO consiste à masquer les zones où le nombre d'exploitations pratiquant le type de production représenté par la carte est réduit ou nul. Comme le montre la figure 9, des régions différentes sont ainsi masquées selon le type de production visé.

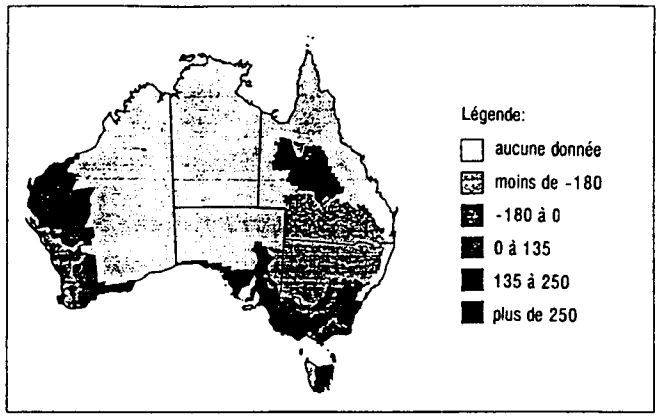

Figure 9. Carte en polygones illustrant la variation probable de la production de laine, de 1991-1992 à 1992-1993, pour les exploitations possédant 100 moutons ou plus en 1991-1992 (kg)

5. DISCUSSION

Dans le présent article, nous avons démontré que lorsque des données d'enquête ont une dimension spatiale, comme c'est le cas pour le AAGIS, les notions de lissage spatial peuvent être utiles à l'analyste. On peut y avoir recours pour modifier les poids et réduire ainsi la variabilité des estimations portant sur les petites régions. Il peut également être utile de procéder au lissage des dimensions spatiales des données avant la cartographie de la fonction de la moyenne spatiale.

Comme le présent article porte sur la cartographie des données, nous avons limité notre propos au lissage des dimensions spatiales. Toutefois, les mêmes méthodes peuvent servir au lissage en fonction d'autres dimensions. Ainsi, si on a des raisons de s'attendre à une forte corrélation sérielle lorsque la population sous-jacente est ordonnée en fonction d'une variable donnée, on pourra alors envisager d'utiliser ces méthodes pour cartographier le «changement» des variables de l'enquête en fonction du changement de cette variable particulière. Il conviendra de noter à ce propos que de telles «cartes» constitueront des estimations non paramétriques des moyennes conditionnelles des variables de l'enquête, compte tenu de cette variable d'«ordonnancement» ou de «lissage». L'analyste devra toutefois tenir compte du grave problème de la dimensionnalité: la taille de l'échantillon effectif diminue radicalement avec chaque nouvelle variable de lissage utilisée avec ces techniques non paramétriques.

Finalement, pour la cartographie des données d'enquête, nous avons utilisé des méthodes d'estimation fondées sur le noyau. Toutefois, on pourrait également recourir au lissage «spline», ou même à des méthodes paramétriques. À notre avis, le choix de la technique de lissage est quelque peu subjectif et dépend du but visé, puisqu'il n'y a pas de raisons objectives définitives pour préférer l'une ou l'autre méthode.

REMERCIEMENTS

Les auteurs tiennent à remercier les examinateurs pour leurs utiles commentaires qui ont grandement amélioré la présentation du présent article.

ANNEXE

Au cours des dernières années, on a établi un certain nombre de propriétés d'optimalité pour les poids du noyau localement linéaires (voir par exemple Wand et Jones (1995) et les articles auxquels ils font références). Nous avons par conséquent comparé les séquences de poids de Nadaraya-Watson (NW) et les séquences localement linéaires (LL) en utilisant des largeurs de bande fixes (LBF) et les largeurs de bande du k-ième plus proche voisin (PPV) avec chaque séquence de poids. Pour chacune de ces combinaisons, nous avons choisi la largeur de bande en utilisant la contre-validation des moindres carrés (CV) ou une méthode spéciale (décrite dans le dernier paragraphe de la présente section) destinée à atténuer l'aspect tacheté des cartes (AATC).

Deux critères nous ont servis à évaluer la performance de chaque méthode. Le premier, l'erreur quadratique moyenne (EQM), est le critère statistique évident pour l'évaluation d'un estimateur biaisé. Le deuxième critère est plus particulier au ABARE. À mesure qu'on produit les estimations en tableaux (par État) et en cartes, l'impression que nous laisse la carte en ce qui concerne la moyenne pour l'État devrait être proche de la valeur indiquée au tableau. Nous avons donc utilisé une somme pondérée du carré des différences observées entre les moyennes pour l'État des données brutes et des données lissées (EB^2). Ce paramètre a également été calculé au niveau régional (RB^2; chaque État contenant de une à neuf régions).

Les données ont été générées pour chaque emplacement en utilisant trois fonctions au degré de lissage différent (mesurées par $\int m''$) et des erreurs de mélange normal. Par exemple,

$$m_1(z) = 6.25 \times 10^4 \times \cos\left(\frac{z_1 - 132.5}{2.25}\right)\cos\left(\frac{z_2 + 27.5}{1.75}\right),$$

où z_1 et z_2 représentent la longitude et la latitude du point z. Les fonctions $m_i(z)$ ont été mises à l'échelle de manière à présenter la même étendue que les valeurs lissées d'une variable d'enquête clé, et les erreurs ont été mises à l'échelle pour présenter la même étendue que les résiduelles de la même variable après le lissage. Les emplacements caractérisés par des résiduelles importantes ont donné des valeurs élevées de la variance, et celles dont les résiduelles étaient faibles ont donné de petites valeurs de la variance. Les résultats de la simulation fondés sur la fonction de lissage sont présentés au tableau 5.

Tableau 5

Comparaison des séquences de poids localement linéaires (LL) et de Nadaraya-Watson (NW), utilisant des largeurs de bande fixes (LBF) et du k-ième plus proche voisin (PPV) choisies par contrevalidation des moindres carrés (CV) et à l'aide du critère décrit ci-après (AATC). Les résultats ont été obtenus à partir de 400 échantillons indépendants avec la fonction moyenne et des erreurs de mélange normales. Les valeurs de l'erreur quadratique moyenne ont été calculées à l'aide de la moyenne de la population finie de $(y - \hat{m}(z))^2$

		EQM × 10^{-7}		RB^2 × 10^{7}		EB^2 × 10^{7}	
		CV	AATC	CV	AATC	CV	AATC
LL	LBF	39.64	93.93	4.44	1.67	1.33	0.39
	PPV	20.50	22.83	2.22	1.35	0.37	0.14
NW	LBF	41.91	52.78	3.29	1.77	0.34	0.17
	PPV	21.77	22.22	3.03	2.33	0.62	0.41

L'utilisation de l'erreur quadratique moyenne en guise de critère d'évaluation de la méthodologie n'a pas donné de résultats cohérents pour les trois fonctions $m_i(z)$. Toutefois, lorsque les valeurs RB^2 ou EB^2 ont été utilisées pour la mesure de la performance, l'estimateur LL avec une largeur de bande du k-ième plus proche voisin, sélectionné à l'aide de

la méthode AATC, a donné des résultats supérieurs d'au moins dix pour cent au résultat des autres méthodes pour chacune des fonctions $m_i(z)$, et a donc été retenu de préférence pour la production des cartes du ABARE.

La méthode de sélection de la largeur de bande visant à réduire l'aspect tacheté d'une carte (AATC) est une mesure du lissage de la carte: elle détermine le degré de similitude de la valeur lissée d'une exploitation quelconque, par rapport à celle de ses voisines. Désignons par $p(i)$ l'estimation du percentile de la variable lissée à la i-ième exploitation. Désignons en outre par S_i l'ensemble des indices des six exploitations les plus proches de cette i-ième exploitation. Cette méthode est fondée sur le calcul de

$$SF(h) = (6n)^{-1} \sum_{\substack{i \\ k \in S_i}} |p(i) - p(k)|$$

Cette valeur est indépendante de l'échelle et diminue monotoniquement avec la réduction de la largeur de bande. La largeur de bande choisie est la plus faible qui présente un taux de réduction de AATC ($< \epsilon$) suffisamment faible. La valeur de ϵ a été choisie subjectivement à la suite d'un examen détaillé des cartes de cinq variables clés pour cinq valeurs de ϵ.

BIBLIOGRAPHIE

BANKIER, M.D., RATHWELL, S., et MAJKOWSKI, M. (1992). Two step generalised least squares estimation in the 1991 Canadian Census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*. Statistics Sweden, Örebro, Octobre 5-7.

BARDSLEY, P., et CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BRECKLING, J., et CHAMBERS, R.L. (1988). *M*-quantiles. *Biometrika*, 75, 761-771.

DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

FULLER, W.A., LOUGHIN, M.M., et BAKER, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186-190.

NEWEY, W.K., et POWELL, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819-847.

RUPPERT, D., et WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, à paraître.

WAND, M.P., et JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā*, série A, 26, 101-116.