Techniques d'enquête, décembre 1990
Vol. 16, n° 2, pp. 267–282
Statistique Canada

Modélisation spatiale et temporelle de données agrégées sur les naissances

DAVID R. BRILLINGER[1]

RÉSUMÉ

À l'aide de graphiques et de cartes de la province de Saskatchewan, nous faisons une analyse des naissances enregistrées en 1986 et 1987 par division de recensement. Nous cherchons à déterminer de quelle manière le nombre des naissances est lié aux périodes de l'année et aux régions géographiques; à cette fin, nous établissons des cartes en courbes de niveau qui décrivent le phénomène des naissances de façon uniforme. Le fait qu'il s'agit de données agrégées pose un problème majeur. En deuxième lieu, nous voulons vérifier dans quelle mesure le modèle normal logarithmique de Poisson peut remplacer, pour des données discrètes, le modèle de régression normal pour variables aléatoires continues. À cette fin, une hiérarchie de modèles pour variables aléatoires discrètes sont ajustés aux observations par la méthode du maximum de vraisemblance; il s'agit du modèle de Poisson ordinaire, du modèle de Poisson avec effet des années et des jours ouvrables et du modèle normal logarithmique de Poisson avec effet des années et des jours ouvrables, l'utilisation de ce dernier étant justifiée par l'absence de covariables importantes dans le processus d'ajustement. Comme nous l'indiquons dans l'article, il s'agit là de résultats provisoires.

MOTS CLÉS: Données agrégées; emprunt d'information; établissement de cartes en courbes de niveau; variation non représentée par la loi de Poisson; analyse à pondération locale; cartes; périodogramme; distribution de Poisson; distribution normale logarithmique de Poisson; effets aléatoires; données géographiques; séries chronologiques; covariables non mesurées.

1. INTRODUCTION

Dans cet article, nous nous intéressons à des données qui ont été enregistrées suivant des périodes de temps et des régions géographiques. Il devrait être facile d'analyser de telles données à cause des possibilités de représentation graphique (par exemple, taux en fonction de la période ou nombre en fonction de la région géographique, comme dans le cas de la représentation graphique de résidus, si souvent utilisée dans l'analyse de régression); dans le cas qui nous occupe toutefois, l'agrégation d'éléments de base soulève des difficultés majeures.

Les données analysées ici concernent essentiellement le nombre quotidien de naissances chez les femmes de 25 à 29 ans pour les années civiles 1986 et 1987 pour chacune des 18 divisions de recensement de la Saskatchewan. Nous nous servons aussi des chiffres de population correspondants, établis lors du recensement de 1986, pour le calcul de taux. Nous avons choisi la Saskatchewan pour cette étude pilote parce que son territoire est modérément étendu et que les limites de ce territoire et des divisions de recensement sont assez régulières. (La seconde raison était importante au début de l'étude parce que nous ne disposions pas de cartes produites par ordinateur.) Nous avons choisi les femmes de 25 à 29 ans parce que c'est le groupe d'âge auquel correspond le plus grand nombre de naissances. Les données nous ont été fournies par Statistique Canada. Elles ont pour caractéristique d'être agrégées, non gaussiennes et non stationnaires dans l'espace et le temps.

Nous cherchons à déterminer de quelle manière le nombre des naissances est lié au temps et à l'espace et, plus particulièrement, à découvrir des régimes de fécondité selon les périodes et les régions et, peut-être, des tendances inédites. L'étude comporte essentiellement deux volets. Nous présentons tout d'abord une analyse à pondération locale de données agrégées; ensuite,

[1] David R. Brillinger, Département de statistique, Université de Californie, Berkeley, CA, 94720, É.-U.

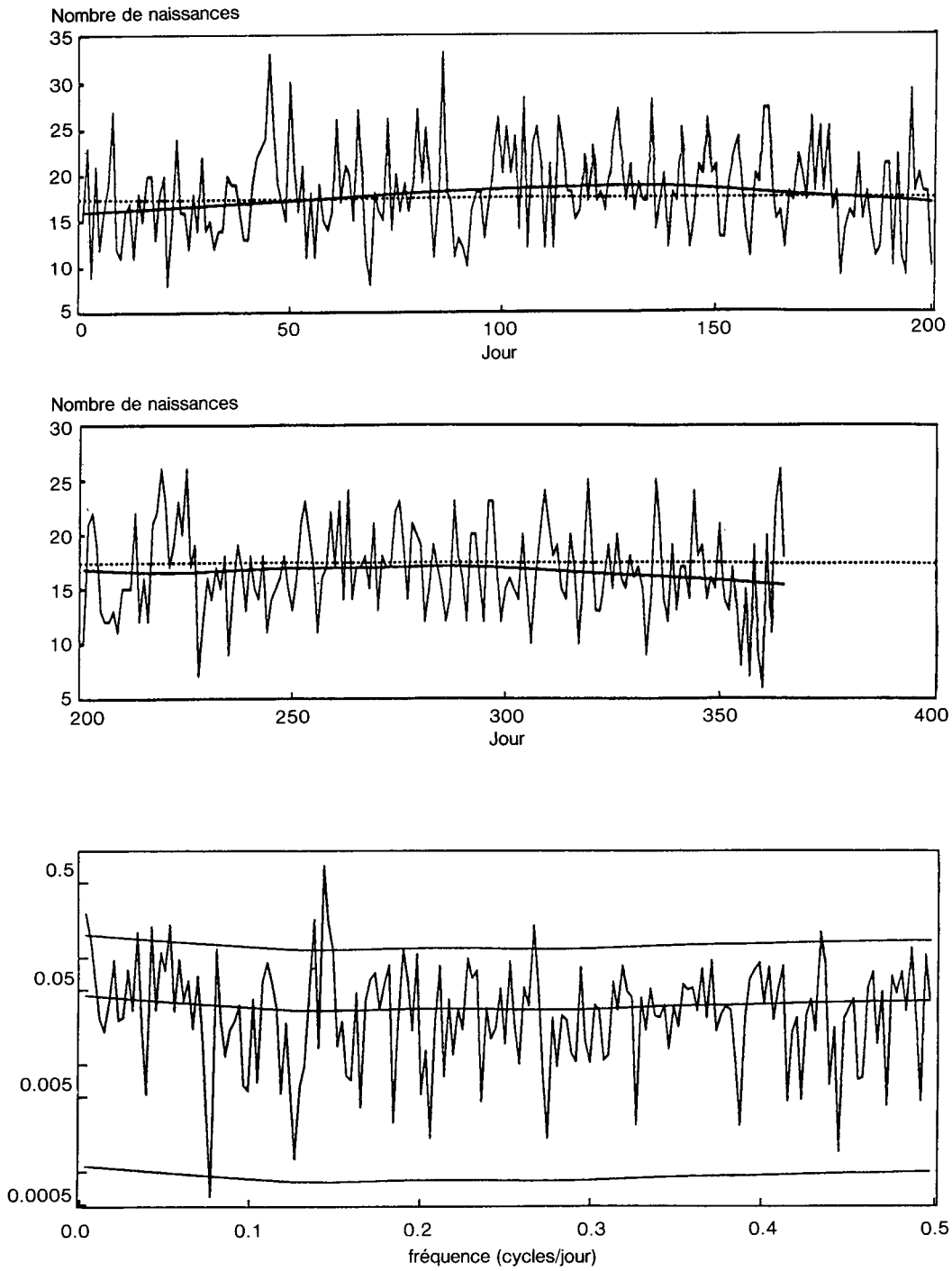

Figure 1. Partie supérieure: Série des naissances enregistrées chez les femmes de 25 à 29 ans en 1986 pour la province de Saskatchewan. Partie inférieure: Périodogramme de la racine carrée du nombre de naissances représenté dans le graphique de la partie supérieure. Les lignes continues représentent des limites de confiance à 95% environ. Le sommet correspond à une période de 7 jours.

nous définissons des modèles avec effets aléatoires, que nous ajustons en vue de traiter la variation non représentée par la loi de Poisson. Le second volet peut être envisagé comme un exercice visant à évaluer la capacité du modèle normal logarithmique de Poisson d'intégrer les covariables non mesurées et les erreurs. L'analyse à pondération locale consiste à construire des poids, $w_i(x,y)$, qui sont censés traduire l'effet de la division de recensement i (un agrégat) sur le lieu géographique défini par les coordonnées (x,y). Les données de la division de recensement étant connues, on applique ensuite ces poids à chacun des termes de la fonction de vraisemblance logarithmique ou des équations d'estimation correspondantes, puis on calcule les valeurs estimées des paramètres.

Nous tenons à préciser qu'il s'agit là d'un rapport provisoire sur des recherches en cours. Par exemple, nous ne prenons pas en considération la structure très détaillée des données et ne produisons aucune mesure de la variance des diverses estimations. Les expressions que nous utilisons pour représenter les poids sont élémentaires et devraient subir des transformations au fur et à mesure des recherches; malgré cela, on peut penser que notre analyse conserve un certain intérêt par son caractère.

Dans un autre article portant sur le même sujet (Brillinger 1990), nous examinons la question sous l'aspect géographique seulement.

2. REPRÉSENTATION CHRONOLOGIQUE DU NOMBRE DE NAISSANCES

Le graphique de la partie supérieure de la figure 1 reproduit le nombre total de naissances enregistrées en Saskatchewan pour chaque jour de 1986. La ligne pointillée représente le nombre moyen de naissances pour 1986. La ligne continue est le résultat d'un lissage accentué de la série et est censée faire ressortir toute tendance. Un examen sommaire de ce graphique ne nous révèle pas de phénomène particulier; cependant, lorsqu'on trace le périodogramme de la racine carrée du nombre de naissances (voir partie inférieure de la figure 1), on constate quelque chose d'intéressant. (La racine carrée sert à rendre la série plus symétrique et plus normale.) Les lignes continues du haut et du bas représentent des limites de confiance marginales à 95% environ pour une série fortement lissée. Un sommet est facilement observable à la fréquence .143 cycle/jour, ce qui correspond à une période de 7 jours. Ce phénomène périodique est fort connu dans l'analyse des données sur les naissances (voir, par exemple, Cohen (1983) et Miyaoka (1989) et les bibliographies qu'ils contiennent). Il s'explique habituellement par le fait que les médecins interviennent et provoquent l'accouchement, surtout les jours de semaine.

3. REPRÉSENTATION GÉOGRAPHIQUE DU NOMBRE DE NAISSANCES

La figure 2 indique, pour chaque division de recensement, le taux annuel de natalité chez les femmes de 25 à 29 ans pour les années 1986 et 1987 combinées. On observe le taux annuel le plus élevé (.208 naissance par femme) dans la partie septentrionale de la province et les deux taux les plus faibles, dans les divisions de recensement où se trouvent les villes de Regina et de Saskatoon.

La figure 3 donne, pour chacune des 18 divisions de recensement, l'écart numérique entre les taux annuels de natalité pour 1987 et 1986. (Précisons que le chiffre de population de 1986 a servi de diviseur dans chaque cas.) Les écarts se situent autour de 0. Il convient toutefois de préciser que ces taux reposent sur des tailles de population très variées.

Dans la section précédente, nous avons constaté l'existence d'un phénomène ayant une période de 7 jours. La figure 4 indique, pour chaque division de recensement, la différence entre le taux de natalité moyen pour les jours de semaine et le taux moyen pour les fins de

semaine (jours de semaine = lundi au vendredi). Le premier est plus élevé que le second dans toutes les divisions sauf une. Ces résultats sont conformes à ceux de diverses autres études et, comme nous le disions dans la section précédente, s'expliquent fort probablement par le fait que les médecins agissent de manière à favoriser les accouchements les jours de semaine (pour ne pas devoir assister à des accouchements les fins de semaine).

Les taux indiqués dans les figures 2, 3 et 4 représentent des valeurs moyennes pour les divisions de recensement.

4. CONTRAINTES DE LA REPRÉSENTATION GÉOGRAPHIQUE DES DONNÉES

Méfiez-vous des cartes qui assignent une valeur moyenne à tout un territoire... (TRADUCTION)

Par ces mots, Tukey (1979) déplore l'utilisation de cartes comme celles représentées par les figures 2, 3 et 4, où une valeur unique est associée à chaque division géographique. En fait, en examinant la figure 2, il est plus logique de penser que le taux de natalité ne varie pas aussi brusquement d'une division de recensement à l'autre. Un des objectifs de notre étude est justement d'établir des cartes en courbes de niveau qui décrivent une variation progressive des taux de natalité. Nous espérons que ces cartes permettront d'en arriver à des descriptions stochastiques générales du phénomène et favoriseront des analyses exploratoires éclairantes.

Dans notre étude, nous nous intéressons aussi à la distribution statistique des effectifs proprement dits. Un modèle stochastique spécial qui s'impose naturellement dans les circonstances est le modèle de Poisson. Or, nous savons, d'après des études antérieures, que les naissances sont liées à de nombreuses variables socio-économiques comme le régime alimentaire, le mode de vie, les conditions atmosphériques, l'environnement, les jours de semaine, les jours fériés, la structure par âge. En outre, la population des divisions de recensement a fluctué autour des chiffres du recensement tout le long de l'année 1986 et 1987 et, finalement, l'âge des femmes visées par l'étude varie de 25 à 29 ans. En résumé, il semble que nous devions utiliser un modèle plus souple que celui de Poisson, c'est-à-dire un modèle qui pourrait tenir compte des covariables absentes. Nous utiliserons donc la distribution normale logarithmique de Poisson dans les circonstances. Par ailleurs, à cause de la présence du paramètre d'écart-type dans cette distribution, nous recourrons à l'"emprunt d'information" en combinant les valeurs d'observations de la manière décrite par Mallows et Tukey (1982). (Nous utilisons le terme "emprunt d'information" plutôt que "estimation empirique de Bayes", que certains préfèrent, parce qu'il a un sens plus large et qu'il est en usage depuis longtemps.) Dean et coll. (1989) est une autre source récente qui traite de la variation non représentée par une loi donnée.

5. ANALYSE À PONDÉRATION LOCALE

Dans le cas de données non agrégées, l'ajustement à pondération locale est une méthode appropriée pour estimer des quantités qui varient graduellement. Supposons que nous avons une variable aléatoire Y avec une distribution de probabilité $p(Y \mid \Theta)$ qui dépend du paramètre de dimension finie Θ. Supposons par ailleurs que nous voulons estimer Θ pour la position définie par les coordonnées (x,y). Supposons enfin que nous connaissons Y_i pour la position (x_i,y_i). Nous allons définir un poids $W_i(x,y)$ qui dépend de la distance entre (x_i,y_i) et (x,y).

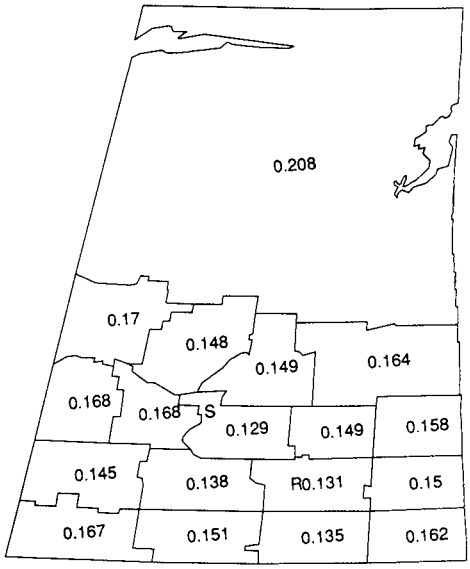

Figure 2. Taux annuel moyen de natalité chez les femmes de 25 à 29 ans pour les années 1986 et 1987, par division de recensement. "R" et "S" indiquent où sont situées les villes de Regina et de Saskatoon respectivement.

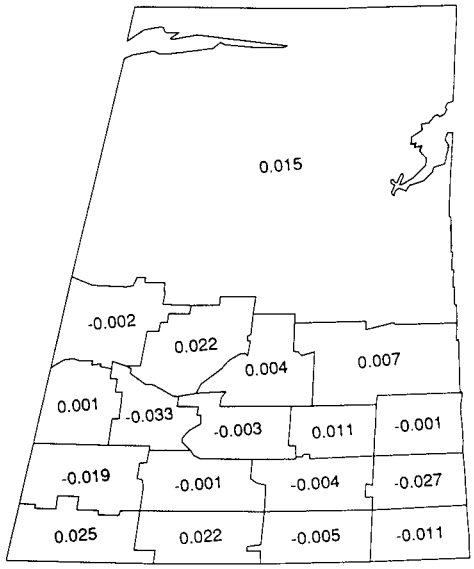

Figure 3. Différence entre le taux pour 1987 et celui pour 1986 (mêmes données que pour la figure 2).

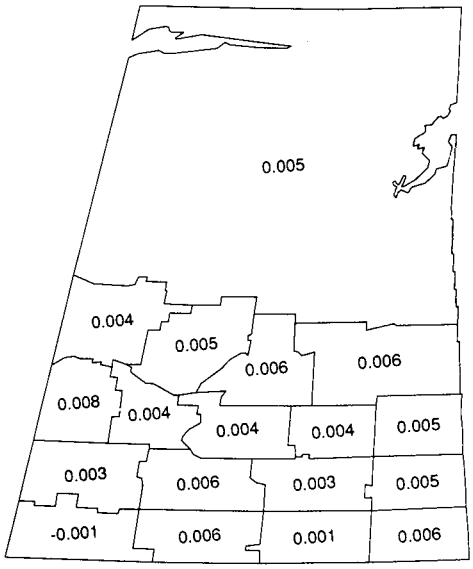

Figure 4. Différence entre le taux moyen pour les jours de semaine et le taux moyen pour les fins de semaine (mêmes données que pour la figure 2).

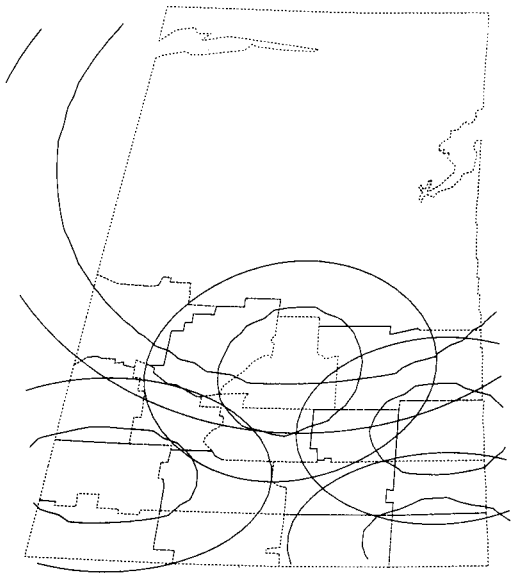

Figure 5. Poids $W_i(x,y)$, utilisés dans les équations (1) ou (2) et calculés au moyen de l'équation (4), pour quatre divisions de recensement. Les divisions ne sont pas toutes représentées pour des raisons de clarté. Les contours à .50 et .99 y sont reproduits.

Il s'agit d'estimer Θ en maximisant la fonction de vraisemblance logarithmique pondérée

$$\sum_i W_i(x,y) \log p(Y_i \mid \Theta) \tag{1}$$

ou (ce qui revient souvent au même) en résolvant le système d'équations d'estimation

$$\sum_i W_i(x,y)\ \Psi(Y_i \mid \hat{\Theta}) = 0 \tag{2}$$

où $\Psi(Y \mid \Theta) = \partial \log p/\partial\Theta$, la fonction de caractérisation.

Afin d'illustrer la technique, prenons un cas simple, celui où Y suit une distribution normale de moyenne μ et de variance σ^2. On obtient la valeur estimée (à pondération locale) de μ à (x,y) en minimisant l'expression

$$\sum_i W_i(x,y)\ [Y_i - \mu]^2$$

ce qui donne

$$\hat{\mu}(x,y) = \sum_i W_i(x,y)\ Y_i \Big/ \sum_i W_i(x,y),$$

expression à première vue intéressante. Soulignons que ces formules sont souvent utilisées en infographie pour l'interpolation de données (voir, par exemple, Franke (1982)).

Parmi les textes traitant ce sujet, mentionnons l'article de Gilchrist (1967) sur l'"actualisation", celui de Pelto et coll. (1968) sur les moindres carrés, celui de Cleveland et Kleiner (1975), qui propose l'utilisation de moyennes mobiles, et celui de Stone (1977), qui porte plus particulièrement sur la régression. Dans son analyse de l'article de Stone, Brillinger (1977) propose l'équation (2) pour une distribution générale et justifie sa proposition en assimilant cette équation à une règle de Bayes. Considérons, en particulier, la fonction de perte

$$L(Y \mid \Theta) = -\log p(Y \mid \Theta).$$

Supposons que nous voulons obtenir la valeur estimée d'un paramètre à $\mathbf{r} = (x,y)$. On peut exprimer le risque de Bayes par l'équation

$$E\{L(Y \mid \Theta_{\mathbf{r}})\} = E\{E\{L(Y \mid \Theta_{\mathbf{r}}) \mid \mathbf{r}\}\}.$$

La règle de Bayes vise à déterminer

$$\min_{\Theta} E\{L(Y \mid \Theta) \mid \mathbf{r}\}.$$

À l'aide des données Y_i, des coordonnées $\mathbf{r}_i$ et de $W_i(\mathbf{r})$, un noyau centré sur $\mathbf{r}_i$, on peut déterminer approximativement l'espérance conditionnelle par la formule

$$E\{\log p(Y \mid \Theta) \mid \mathbf{r}\} \approx \sum_i W_i(\mathbf{r}) \log p(Y_i \mid \Theta)$$

ce qui nous renvoit à l'expression (1).

Tibshirani et Hastie (1987) ont élaboré une méthode d'estimation à équipondération locale fondée sur le principe de la vraisemblance. Cleveland et Devlin (1988) font un traitement très détaillé de la méthode des moindres carrés. Pour sa part, Staniswalis (1989) étudie et applique le cas p général. Les avantages de la technique à pondération locale sont les suivants: aucune hypothèse "cachée" sur la distribution du modèle, mise en évidence des cas de non-additivité, variantes pour la résistance et l'influence, additivité simple des observations et aucune inversion de matrice (comme l'exige le krigeage par exemple).

Les données qui nous intéressent ici consistent essentiellement en des totaux pour des divisions de recensement. Nous ne pouvons donc pas utiliser directement la méthode décrite dans la section précédente. Il s'agit ici de déterminer des poids $w_i(x,y)$ qui traduisent convenablement l'effet de la division de recensement i sur la position (x,y). Supposons que $|R_i|$ désigne la superficie de la division de recensement i. Alors, la fonction de poids élémentaire est

$$w_i(x,y) = \frac{1}{|R_i|} \quad \text{pour } (x,y) \text{ dans } R_i$$

et 0 dans le cas contraire. Dans cette étude, nous allons utiliser des fonctions fondamentales comme

$$w_i(x,y) = \frac{1}{|R_i|} \int_{R_i} W(x-u, y-v)\,du\,dv \tag{3}$$

où $W(\cdot)$ est un noyau qui convient à des données non agrégées comme celles analysées, par exemple, dans Cleveland et Devlin (1988). Pour justifier l'utilisation de l'équation (3), nous pouvons considérer un processus ponctuel de Poisson (voir Annexe II). Les estimations seront calculées à l'aide des formules (1) ou (2), W_i étant remplacé par w_i.

Dans cette étude initiale, les poids particuliers utilisés à $\mathbf{r} = (x,y)$ sont

$$w_i(\mathbf{r}) = \exp\{-(1-\rho)^2 \|\mathbf{r} - \mathbf{r}_i\|^2 / 2\tau^2\} \tag{4}$$

à l'extérieur de l'ellipse $(\mathbf{r}_0 - \bar{\mathbf{r}}_i)\mathbf{S}_i^{-1}(\mathbf{r}_0 - \mathbf{r}_i)' = d_0^2 = 5.991$ et 1 à l'intérieur. Dans l'équation (4), $\|\mathbf{r}\|^2 = x^2 + y^2$, $\rho = d_0 / \sqrt{(\mathbf{r} - \bar{\mathbf{r}}_i)\mathbf{S}_i^{-1}(\mathbf{r} - \bar{\mathbf{r}}_i)'}$ et $\tau = .025$, où $\bar{\mathbf{r}}_i = E\,U_i$ et $\mathbf{S}_i = \operatorname{var} U_i$, U_i étant une variable aléatoire distribuée uniformément dans R_i. Le paramètre ρ est défini de manière que la fonction de poids soit continue. En clair, cela signifie que les divisions de recensement sont représentées par des ellipses ayant la même moyenne et la même matrice de variances-covariances. (Les valeurs exactes ont été déterminées après quelques essais pour faire en sorte que la surface de la première ellipse équivaille à environ 95% de la superficie de la division de recensement.) Nous aurions pu choisir d'autres figures que l'ellipse - un rectangle par exemple -mais la recherche sur ce sujet est encore jeune et il est permis de croire que les études ultérieures utiliseront des poids comme celui défini en (3).

La figure 5 illustre les contours à .50 et à .99 des poids $w_i(x,y)$ pour plusieurs divisions de recensement. On remarque que ces contours suivent la forme générale des divisions. La forme irrégulière de certains contours est attribuable au caractère discret de la grille de 40 × 40 utilisée dans les calculs.

Tobler (1979) et Dyn et Wahba (1982) décrivent d'autres fonctions de poids construites dans des circonstances semblables. La méthode que nous venons d'exposer présente quelques-uns des avantages de la technique à pondération locale: additivité des termes des expressions (1) et (2) et absence d'interaction; aucune inversion de matrice requise; et intégration facile d'un mécanisme de résistance aux valeurs aberrantes.

Cliff et Ord (1975) (section 5.1) cherchent à mesurer l'influence que des comtés exercent l'un sur l'autre. Notre article a plutôt pour but d'analyser l'influence d'un "comté" sur un point géographique en particulier. Peut-être le poids, qui rend compte de l'influence, devrait-il dépendre de covariables (par ex.: la population de comté)?

6. AJUSTEMENT DU MODÈLE DE POISSON

Dans cette analyse, nous nous intéressons uniquement aux femmes de 25 à 29 ans et aux naissances enregistrées chez ce groupe de femmes. Désignons par $i = 1, \ldots, 18$ la division de recensement et par N_i le nombre de femmes recensées dans la division i (il s'agit des chiffres du recensement du 3 juin 1986). Soit B_i le nombre total de naissances enregistrées chez les femmes de 25 à 29 ans en 1986 et 1987.

Supposons que la distribution de probabilité $p(\cdot)$ de B_i (section 5) est une distribution de Poisson de moyenne $2N_i\mu$ (le chiffre 2 signifiant que le paramètre μ représente un taux annuel de natalité). Cette hypothèse repose sur l'idée que les anniversaires de naissance sont aléatoires (voir Brillinger 1986).

Étant donné l'hypothèse du modèle de Poisson, la valeur estimée (à pondération locale) du taux annuel de natalité à la position (x,y) est

$$\hat{\mu}(x,y) = \sum_i w_i(x,y)\, B_i \Big/ 2 \sum_i w_i(x,y) N_i. \tag{5}$$

Ces valeurs sont calculées pour (x,y) sur une grille de 40 par 40 et le tracé de contours correspondant est reproduit dans la figure 6. On observe une progression lente de la valeur des contours. Le taux varie de .14 à .20, les valeurs les plus élevées étant observées dans la partie septentrionale de la province et les moins élevées étant concentrées dans les régions les plus urbanisées.

Nous avons mentionné plus haut que les données à l'étude avaient une dimension temporelle importante. Les modèles doivent tenir compte de cette réalité. En particulier, ils doivent pouvoir décrire la période hebdomadaire qui caractérise ces données, ainsi que les tendances démographiques possibles. Voici donc un modèle qui mérite d'être considéré. Soit j une variable indicatrice qui prend la valeur 1 si c'est pour un jour de semaine et la valeur 2 si c'est pour un jour non ouvrable. Soit k une seconde variable indicatrice qui prend la valeur 1 pour 1986 et la valeur 2 pour 1987. Désignons par B_{ijk} le nombre de naissances dans la division de recensement i. Supposons que B_{ijk}, étant donné N_i, suit une distribution de Poisson de moyenne $N_i \exp\{\alpha + \beta_j + \gamma_k\}$. β_j représente l'effet des jours ouvrables et γ_k, l'effet d'année, et nous supposons que $\beta_1 + \beta_2, \gamma_1 + \gamma_2 = 0$ pour que le modèle soit identifiable. S'il n'y a pas d'effet des jours ouvrables, alors, $\beta_1, \beta_2 = 0$. S'il n'y a pas d'effet d'année, alors $\gamma_1, \gamma_2 = 0$. Par ailleurs, en utilisant la méthode d'estimation à pondération locale décrite dans la section 5, on peut estimer α, β et γ et comme des fonctions de la position (x,y). (Pour simplifier les calculs, nous avons utilisé uniquement les 364 ($= 7 \times 52$) premiers jours de chaque année.)

La figure 7 donne la valeur estimée $\exp\{\hat{\alpha}(x,y)\}$ du taux annuel de natalité. Il est intéressant de constater que, par rapport à la figure précédente (modèle de Poisson ordinaire), les contours sont un peu plus éloignés des régions urbaines. La figure 8 illustre l'effet estimé des jours ouvrables, $\hat{\beta}_1(x,y)$. Dans ce cas, on observe une poussée dans l'est de la province. Cette figure tranche considérablement avec la figure 4, où l'on se contente de reproduire de simples écarts par région. Le trait distinctif de la figure 8 est qu'elle reflète les différences de population. Les valeurs de $\hat{\beta}$ vont de .08 à .13 tandis que celles de $\hat{\alpha}$ vont de -2.1 à -1.6. La figure 9 donne l'effet d'année estimé, $\hat{\gamma}_1(x,y)$. Les valeurs de cet effet vont de $-.03$ à .03. En termes numériques, l'effet des jours ouvrables est plus grand que l'effet d'année.

L'analyse que nous venons de faire donne à penser que des variables fondamentales peuvent influer sur les taux de natalité et que nous devons en tenir compte dans la modélisation et l'analyse.

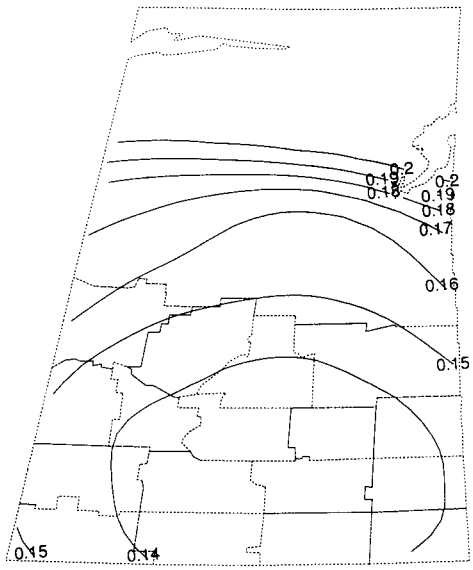

Figure 6. Représentation graphique de l'équation (5), où les poids sont ceux définis par l'équation (4), B_i est le nombre de naissances enregistrées dans la division de recensement i en 1986 et 1987, et N_i est l'effectif des femmes de 25 à 29 ans dans la division i pour la même période.

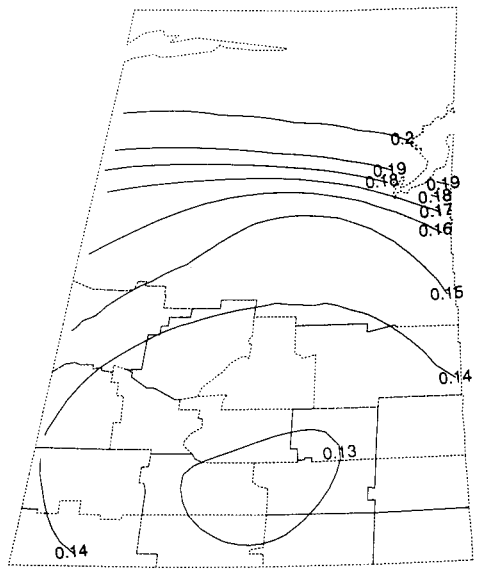

Figure 7. Taux de natalité estimé, $\exp\{\hat{\alpha}\}$ calculé au moyen de l'ajustement à pondération locale suivant l'hypothèse que le nombre des naissances, B, étant donné la population visée N, suit une distribution de Poisson de moyenne $N \exp\{\alpha \pm \beta \pm \gamma\}$, où le premier signe $\pm$ se rapporte à l'effet des jours ouvrables (+ pour un jour de semaine; − pour la fin de semaine) et le second, à l'effet d'année (+ pour 1986; − pour 1987).

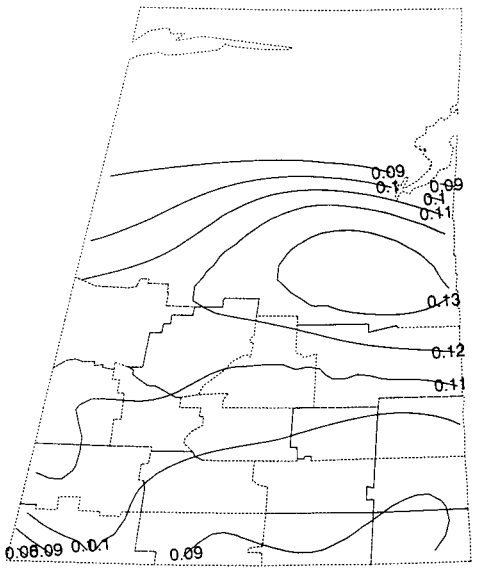

Figure 8. Représentation graphique de l'effet estimé des jours ouvrables, $\hat{\beta}(x,y)$, établie suivant la figure 7.

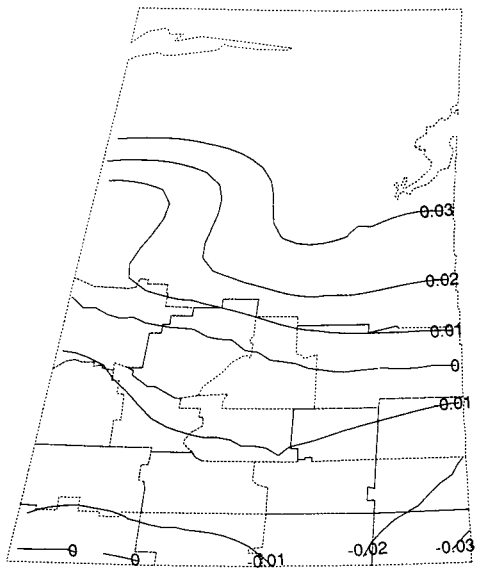

Figure 9. Représentation graphique de l'effet d'année estimé, $\hat{\gamma}(x,y)$, établie suivant la figure 7.

7. AJUSTEMENT DU MODÈLE NORMAL LOGARITHMIQUE DE POISSON

Étant donné une variable explicative multidimensionnelle $\mathbf{x}$, un modèle de Poisson de moyenne $N \exp\{\mathbf{x}\Theta\}$ pour B pourrait expliquer assez bien les données. Comme variables explicatives, pensons au régime alimentaire, au mode de vie, aux conditions atmosphériques, à l'environnement, aux jours fériés, au changement démographique, à la structure par âge, aux caprices des délimitations. Comme nous ne connaissons rien de ces variables dans les circonstances, nous allons supposer que les variables absentes du modèle sont représentées collectivement par une variable d'erreur. Nous allons aussi supposer que, étant donné ϵ, la variable aléatoire B suit une distribution de Poisson de moyenne $N\mu \exp\{\epsilon\}$ et que ϵ suit une distribution normale de moyenne 0 et de variance σ^2. En ce qui concerne ce modèle, on dira que B suit une distribution normale logarithmique de Poisson. Pour plus de renseignements sur cette distribution, veuillez vous référer à Shaban (1988). Il arrive que ϵ découle directement du contexte (voir Brillinger et Preisler (1983) pour un exemple) mais dans le cas qui nous occupe, nous avons simplement supposé son existence.

Un inconvénient majeur du modèle normal logarithmique de Poisson est qu'il n'existe pas d'expression en forme analytique pour la fonction de probabilité. Cependant, le modèle se prête très bien à l'introduction d'effets et au traitement de variables manquantes. Les travaux de Bock et Lieberman (1970), de Pierce et Sands (1975) et de Hinde (1982) nous indiquent que l'on peut recourir à la quadrature numérique. Nous pouvons exprimer la fonction de probabilité par la formule

$$p(Y) = \frac{1}{Y!}\int (ve^{\sigma z})^{Y} \exp\{-ve^{\sigma z}\}\,\phi(z)dz$$

où ϕ est la densité normale standard, Y correspond à B et v correspond à $N\mu$. Pour faire une analyse de données, on remplace l'intégrale par un nombre fini de termes comprenant des noeuds, z_l, et des poids, w_l,

$$p(Y) \approx \frac{1}{Y!}\sum_{l=1}^{L} (ve^{\sigma z_l})^{Y} \exp\{-ve^{\sigma z_l}\}\, w_l.$$

Pour une liste de noeuds et de poids, voir, par exemple, Abramowitz et Stegun (1964).

Les figures 10, 11, 12 et 13 donnent les résultats de l'ajustement du modèle normal logarithmique de Poisson (y compris les versions avec effet des jours ouvrables et effet d'année) fait à l'aide de $L = 5$ noeuds. Le modèle suppose que B_{ijk}, étant donné N_i et Z, suit une distribution de Poisson de moyenne

$$N_i \exp\{\alpha + \beta_j + \gamma_k + \sigma Z\}$$

Z étant un écart normal standard, et suppose de plus que les différents Z sont indépendants. Dans ce modèle, i désigne la division de recensement, j indique s'il s'agit d'un jour de semaine ou non et k désigne l'année. La figure 10, qui représente un tracé de contours pour $\exp\{\hat{\alpha}(x,y)\}$, montre des cercles vaguement concentriques autour des régions urbaines, comme dans la figure 7. La forme irrégulière du tracé donne à penser que le processus d'estimation pourrait, à une occasion, avoir convergé vers un extrémum local. Les figures 11 et 12 donnent les tracés de contours pour $\hat{\beta}(x,y)$ et $\hat{\gamma}(x,y)$ respectivement. Là encore, on a l'impression qu'il y a eu convergence vers un extrémum local. La figure 13, qui représente un tracé de contours pour $\hat{\sigma}(x,y)$, n'est pas facile à décrire. Elle donne à croire que la valeur estimée $\hat{\sigma}$ est passablement variable. Cette valeur se situe autour de 0.1 et est donc comparable à l'effet des jours ouvrables défini dans la section 6.

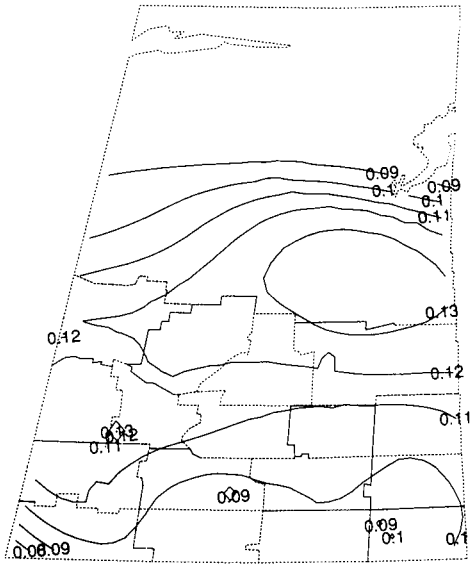

Figure 10. Tracé comparable à celui de la figure 7, sauf qu'un terme d'erreur normal a été ajouté au prédicteur linéaire.

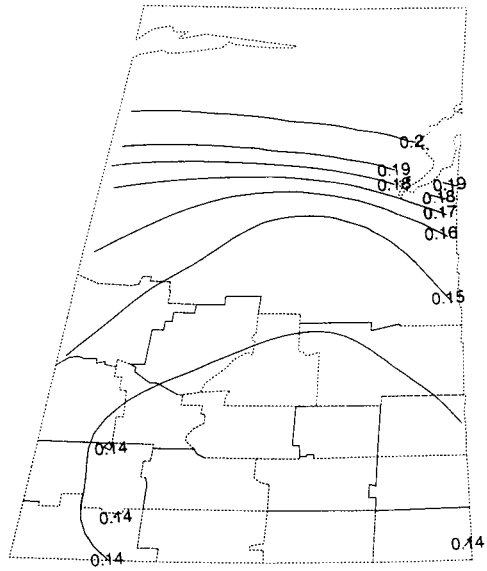

Figure 11. Tracé comparable à celui de la figure 8, sauf que (comme dans la figure 10) un terme d'erreur normal a été ajouté au prédicteur linéaire.

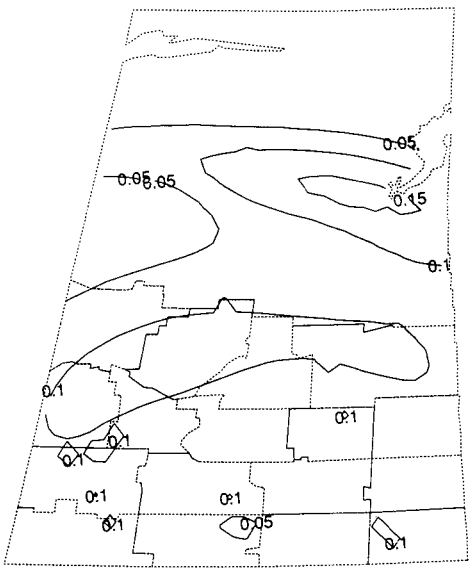

Figure 12. Tracé comparable à celui de la figure 9, sauf que (comme dans la figure 10) un terme d'erreur normal a été ajouté au prédicteur linéaire.

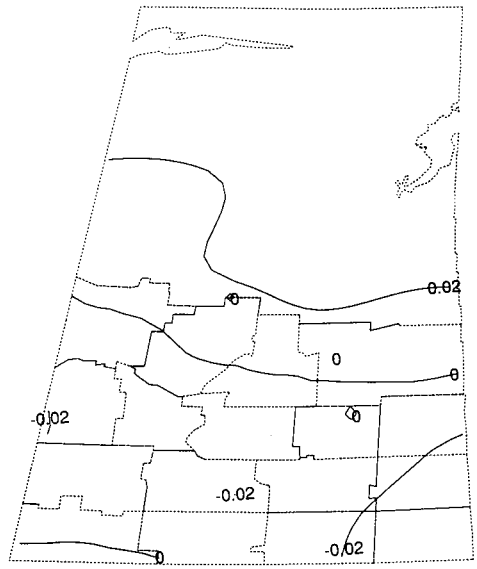

Figure 13. Erreur type estimée, $\hat{\sigma}(x,y)$, du terme normal qui a été ajouté au prédicteur linéaire.

Toutes les études qui traitent l'estimation par le modèle normal logarithmique de Poisson et dont nous connaissons l'existence impliquent une forme quelconque d'approximation. Clayton et Kaldor (1987), par exemple, utilisent une équation quadratique en guise d'approximation pour la fonction de vraisemblance logarithmique conditionnelle de Poisson, et Aitchison et Ho (1989) utilisent à leur tour l'intégration numérique, non sans avoir transformé les paramètres au préalable. Crouch et Spiegelman (1990) ont proposé récemment une nouvelle forme d'approximation, dont on n'a pas encore étudié l'efficacité par rapport au modèle normal logarithmique de Poisson.

8. ANALYSE

L'analyse à pondération locale et les modèles avec effets aléatoires semblent être deux moyens de résoudre avec souplesse toute une série de problèmes ayant trait à des données régionales. Les termes d'effet aléatoire ont deux fonctions importantes: tenir compte des effets manquants et "emprunter" de l'information de manière à produire de meilleures estimations pour les principaux paramètres. En ce qui concerne le modèle de Poisson ordinaire, les totaux élémentaires sont efficaces mais, dans le cas qui nous occupe, il existe une variation non représentée par la loi de Poisson à cause des variables manquantes.

La méthode exige beaucoup de temps d'ordinateur parce que l'intégration numérique et l'estimation par le maximum de vraisemblance sont exécutées à de nombreuses positions sur une grille; toutefois, les opérations se sont bien déroulées sur le système Sun 3/50 de Berkeley.

Beaucoup de recherches restent à faire; mentionnons au passage des sujets comme l'évaluation de l'ajustement, le calcul et la présentation de la variabilité, le choix de la fonction de poids (notamment le choix de τ dans l'équation (4)), les analyses pour d'autres groupes d'âge et d'autres provinces et la définition d'asymptotes appropriées. Il faudra aussi chercher à comprendre pourquoi, avec des valeurs initiales proches de la valeur réelle, le processus d'optimisation convergeait parfois vers des valeurs estimées assez éloignées de la valeur réelle. L'avantage de notre étude est qu'elle renferme une quantité d'autres données qui peuvent servir au fur et à mesure des recherches. En examinant la figure 6 et les suivantes, on remarque une faiblesse importante de la technique: un trop grand niveau de détail dans la partie septentrionale de la province.

Parmi les articles récents portant sur l'analyse des données démographiques, notons ceux de Cressie et Read (1989), de Clayton et Kaldor (1987), de Tsutakawa (1988) et de Manton et coll. (1989). Contrairement à la présente étude, ces articles n'ont pas pour objet d'étudier la question des surfaces lisses.

Il est amusant de savoir que l'existence d'un phénomène d'une période de sept jours nous a amenés à découvrir très tôt qu'il y avait eu confusion au sujet de la série de données à utiliser. En effet, lorsqu'est venu le moment de déterminer, à l'aide de la série initiale de données, les jours où il y avait le moins de naissances, nous avons constaté qu'il s'agissait, selon toute vraisemblance, du vendredi et du samedi parce que nous avions en mains les données de 1987 et non celles de 1986, comme ç'aurait dû être le cas.

Une fois les analyses terminées, nous avons appris que le nombre de naissances avait été établi en fonction des divisions de recensement de 1981 tandis que les chiffres de population, eux, l'avaient été en fonction des divisions de 1986. Heureusement, les limites des divisions ont peu changé entre les deux recensements mais cette méprise offre une raison de plus pour réclamer une méthode qui puisse tenir compte de la variation non représentée par une loi donnée.

9. ADDENDA

Dans cet article, nous avons notamment considéré l'inclusion d'un terme d'erreur (ϵ) dans le modèle afin de représenter les covariables pertinentes qui étaient absentes du modèle. Cela nous a amenés à utiliser la distribution normale logarithmique de Poisson. Tukey (1990) construit un "indice d'urbanisation" pour les divisions de recensement. Cet indice repose sur les chiffres de population des trois plus grandes agglomérations de la division. Les valeurs de l'indice, x_i, sont reproduites dans la figure 14; nous constatons que les valeurs les moins élevées correspondent aux divisions de recensement qui renferment les villes de Regina et de Saskatoon.

Le tableau ci-dessous contient les résultats de l'ajustement de modèles de Poisson pour B_{ijk}, étant donné N_i, au moyen du logiciel GLIM; ces modèles sont: i) $N_i \exp\{\alpha + \beta_j + \gamma_k\}$, ii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta x_i\}$ et iii) $N_i \exp\{\alpha + \beta_j + \gamma_k + \delta_1 x_i + \delta_2 x_i^2\}$.

Variables	Somme des carrés des écarts	d.l.	valeur p
jours ouvrables, année	227.3	69	
+ urbanisation	86.69	68	
+ urbanisation**2	83.13	67	.088

Grâce à l'introduction de la variable d'urbanisation, x_i, nous pouvons maintenant nous accommoder d'un modèle de Poisson ordinaire.

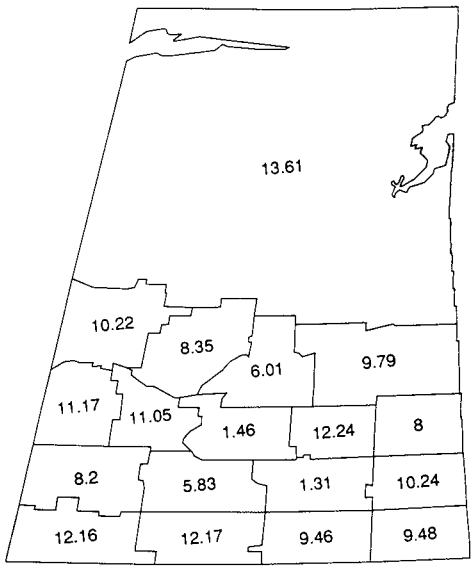

Figure 14. Valeurs de l'indice d'urbanisation de Tukey.

Enfin, l'arbitre a exprimé des commentaires qui mettent très bien en relief les hypothèses et les limites de cette étude. Nous poursuivons notre recherche en vue de corriger les faiblesses qu'il signale. Nous avons jugé raisonnable ici de reproduire textuellement ses commentaires.

"Le choix des poids est trop *spécifique* et exige une plus grande réflexion. Si nous avions, par exemple, deux divisions de même superficie mais dont les populations respectives N_i sont très différentes, devrions-nous leur attribuer le même poids? Tout dépend de ce qu'on juge le plus important entre la superficie ou la densité de population. En optant pour la seconde, nous pourrions faire abstraction du niveau de détail indu que l'on trouve dans la partie septentrionale de la province."

"Les N_i posent des embûches que l'auteur semble ne pas ignorer; néanmoins, nous tenons à mettre en garde davantage le lecteur. Il serait bon d'avoir des mesures approximatives de la variance (dans la section 1, l'auteur indique qu'il n'en produit pas). On ne peut pas vraiment interpréter la figure 3 puisque les valeurs positives ou négatives peuvent être attribuables à des fluctuations aléatoires autour de zéro. Les contours de la figure 6 sont calculés avec un degré de précision très variable et ne sont pas comparables à certains égards. Enfin, pour ce qui a trait à l'estimation de α, β et γ (section 6), il serait tentant (mais imprudent) de supposer que ces valeurs sont significatives."

"Toutes les variables aléatoires mentionnées dans cet article sont supposées indépendantes. Une autre façon de justifier les modèles pondérés est de supposer l'existence d'une distribution multidimensionnelle, où la moyenne conditionnelle à (x,y), compte tenu des données relatives aux positions voisines, est une combinaison pondérée de ces données. On observe alors un lien de dépendance dans la distribution conjointe."

REMERCIEMENTS

L'auteur tient à exprimer sa reconnaissance à G. Brackstone, R. Gussela, R. Raby, B. Sander, P. Spector, M. Subhani et R. Villani pour lui avoir procuré les données et les cartes et l'avoir assisté dans la géométrie computationnelle et le calcul en parallèle. John Tukey, Rob Tibshirani et l'arbitre ont exprimé des commentaires très utiles sur la première version. Cette étude a été rendue possible grâce à une subvention de la National Science Foundation (n° DMS-8900613).

ANNEXE I

Cette annexe contient quelques détails sur les calculs. Les limites de la province et des divisions de recensement forment des polygones. Pour calculer les poids $w_i(x,y)$, il a fallu utiliser un algorithme qui vérifiait si un point donné se trouvait dans un polygone donné. Pour calculer la moyenne et la variance d'un point aléatoire à l'intérieur d'un polygone donné, il a fallu recourir à un algorithme par lequel le polygone était divisé en triangles. Ces algorithmes sont analysés notamment dans Preparata et Shamos (1985). La fonction de vraisemblance approximative a été maximisée au moyen du programme FORTRAN va09a de Harwell. Pour le calcul en parallèle, nous avons décomposé la grille de 40 par 40 en 20 segments disjoints, puis avons effectué les calculs avec 20 postes de travail différents. Comme dans Brillinger et Preisler (1983), nous avons introduit des facteurs dans la fonction de vraisemblance afin de stabiliser les calculs. Miyaoka (1989) a observé que les calculs peuvent être sensibles au nombre de noeuds utilisés. Dans le cas qui nous occupe, nous avons augmenté le nombre de noeuds jusqu'à ce qu'il n'y ait plus de changement perceptible dans les résultats. Enfin, il y a aussi la question du choix des valeurs initiales. Dans cette étude, nous avons eu recours à la méthode des estimateurs de moments; toutefois, ceux-ci sont peut-être trop peu efficaces.

ANNEXE II

Pour plus de simplicité, considérons le cas d'un processus ponctuel $\{x_j\}$ avec une fonction de taux v sur la ligne. La fonction de vraisemblance logarithmique pondérée (localement) pour un processus de Poisson est, à une constante près,

$$\sum_j W(x - x_j) \log v(x_j) - \int W(x - u) v(u) du.$$

Par conséquent, l'estimateur (à pondération locale) du taux est

$$\hat{v}(x) = \sum_j W(x - x_j) \Big/ \int W(x - u) du,$$

qui est la forme habituelle. Supposons maintenant que la ligne est découpée en intervalles R_i et que l'effectif connu est $N(R_i)$. On cherche à calculer

$$\sum_{x_j \in R_i} W(x - x_j).$$

Si on remplace l'expression ci-dessus par une mesure qui en est une approximation, $\Theta N(R_i)$, alors par la méthode des moments, on obtient

$$\Theta = \int_{R_i} W(x - u) du \Big/ | R_i |$$

ce qui nous amène à l'équation (3).

BIBLIOGRAPHIE

ABRAMOWITZ, M., et STEGUN, I.A. (1964). *Handbook of Mathematical Functions*. National Bureau of Standards, Washington.

AITCHISON, J., et HO, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.

BOCK, R.D., et LIEBERMAN, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, 35, 179-197.

BRILLINGER, D.R. (1977). Discussion of Stone (1977). *The Annals of Statistics*, 5, 622-623.

BRILLINGER, D.R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42, 693-734.

BRILLINGER, D.R. (1990). Mapping aggregate birth data. Rapport technique, Statistics Department, University of California, Berkeley.

BRILLINGER, D.R., et PREISLER, H.K. (1983). Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series and Multivariate Statistics*, 31-65. New York: Academic Press.

CLAYTON, D., et KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

CLEVELAND, W.S., et DEVLIN, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

CLEVELAND, W.S., et KLEINER, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics*, 17, 447-454.

CLIFF, A.D., et ORD, J.K. (1975). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society*, 37, 297-348.

COHEN, A. (1983). Seasonal daily effect on the number of births in Israel. *Applied Statistics*, 32, 228-235.

CRESSIE, N., et READ, T.R.C. (1989). Spatial data analysis of regional counts. *Biometrical Journal*, 31, 699-719.

CROUCH, E.A.C., et SPIEGELMAN, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t) \exp\{-t^2\} dt$: application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464-469.

DEAN, C., LAWLESS, J.F., et WILLMOT, G.E. (1989). A mixed Poisson-inverse-Gaussian regression model. *La Revue Canadienne de Statistique*, 17, 171-182.

DYN, N., et WAHBA, G. (1982). On the estimation of functions of several variables from aggregated data. *SIAM Journal on Mathematical Analysis*, 13, 134-152.

FRANKE, R. (1982). Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38, 181-200.

GILCHRIST, W.G. (1967). Methods of estimation involving discounting. *Journal of the Royal Statistical Society*, 29, 355-369.

HINDE, J. (1982). Compound regression models. GLIM 82 (éd. R. Gilchrist), 109-121. *Lecture Notes in Statistics*, 14. New York: Springer-Verlag.

MALLOWS, C., et TUKEY, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects. *Some Recent Advances in Statistics* (éds. Tiago de Oliveira, J. et coll.), 111-172. London: Academic.

MANTON, K.G., WOODBURY, M.A., STALLARD, E., RIGGAN, W.B., CREASON, J.P., et PELOM, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U. S. cancer mortality rates. *Journal of the American Statistical Association*, 84, 637-650.

MIYAOKA, E. (1989). Application of mixed Poisson-process models to some Canadian birth data. *La Revue Canadienne de Statistique*, 17, 123-140.

PELTO, C.R., ELKINS, T.A., et BOYD, H.A. (1968). Automatic contouring of irregularly spaced data. *Geophysics*, 33, 424-430.

PIERCE, D.A., et SANDS, B.R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Department, Oregon State University.

PREPARATA, F.P., et SHAMOS, I. (1985). *Computational Geometry*. New York: Springer-Verlag.

SHABAN, S.A. (1988). Poisson-lognormal distributions. *Lognormal Distributions*, 195-210 (éds. E.L. Crow et K. Shimizu). New York: Marcel Dekker.

STANISWALIS, J.G. (1989), The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, 276-283.

STONE, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 5, 595-620.

TIBSHIRANI, R., et HASTIE, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

TOBLER, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Assocation*, 74, 519-536.

TSUTAKAWA, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 83, 37-42.

TUKEY, J.W. (1979). Statistical mapping: what should not be plotted. Proc. 1976 Workshop on Automated Cartography. DHEW Publication No. (PHS) 79-1254, 18-26. Dans The Collected Works of J.W. Tukey, Vol. 5 (1988), (éd. W.S. Cleveland). Pacific Grove: Wadsworth.

TUKEY, J.W. (1990). Graphical diplays of: Are the (x,y) pairs compatible with a linear dependence? Technical Report No. 301, Princeton University.