Article

« L'autocorrélation spatiale et les données de santé : une étude préliminaire »

Diana C. Bouchard

Cahiers de géographie du Québec, vol. 20, n° 51, 1976, p. 521-538.

CAHIER DE GÉOGRAPHIE DE QUÉBEC,
Vol. 20, numéro 51, décembre 1976, 521-538

L'AUTOCORRÉLATION SPATIALE ET LES DONNÉES DE SANTÉ: ÉTUDE PRÉLIMINAIRE

par

Diana C. BOUCHARD *

Département de géographie, université McGill, Montréal, H3C 3G1

INTRODUCTION ET ÉLÉMENTS MATHÉMATIQUES

L'analyse de données réparties dans l'espace est fondée sur l'hypothèse de l'existence virtuelle de distributions régulières. Malheureusement, la plupart des techniques actuelles destinées à découvrir de telles distributions ne tiennent pas compte de leur contenu spatial et traitent les observations comme une série unidimensionnelle de nombres. Même les méthodes qui prennent explicitement en considération des variables à caractère spatial, telles que la distance ou le degré d'agglomération, évitent généralement de spécifier la relation entre les valeurs observées et leur position dans l'espace. Cette relation s'appelle l'autocorrélation spatiale.

Le problème général de l'autocorrélation spatiale peut être formulé ainsi : étant donné n unités d'observation, chacune avec une valeur x_i, à quel point la variation dans les x_i est-elle due à la contiguïté dans l'espace des n unités d'observation ? Le fait que ce problème se situe, par définition même, en deux dimensions, le distingue nettement de l'analyse d'autocorrélation des séries unidimentionnelles, qui a été élaborée par les économétriciens par le biais de l'analyse des séries temporelles et par les mathématiciens sous forme d'un ensemble de techniques destinées à analyser les erreurs résiduelles d'un modèle de régression.

Tandis que les économétriciens emploient généralement l'analyse d'autocorrélation pour examiner et expliquer des tendances cycliques ou à long terme dans les données économiques, les mathématiciens s'en servent plutôt pour vérifier la validité ou juger de l'exactitude d'un modèle de régression. Si une succession d'erreurs résiduelles manifeste une corrélation différente de zéro, il est fort probable que l'hypothèse fondamentale de l'indépendance de ces erreurs résiduelles est compromise et donc que la plus grande partie de l'inférence statistique classique concernant la régression est mise en doute. L'autocorrélation dans une série d'erreurs résiduelles peut avoir plusieurs causes, telles que l'omission de variables,

* L'auteur remercie le Service de la recherche universitaire et les responsables des programmes FCAC du ministère de l'Éducation du gouvernement du Québec pour avoir soutenu le travail de recherche qui a donné lieu à ce projet, Andrée Deveault qui a relu le manuscrit et contribué à son amélioration, et le professeur Bryan Massam qui a prodigué ses encouragements tout au long de ce travail.

une définition inexacte du modèle de régression, ou la présence de tendances cycliques dans les données. Le modèle simple d'autorégression du premier ordre, où chaque terme n'est influencé que par celui qui le précède immédiatement (Johnston, 1972, p. 244), peut s'exprimer ainsi :

$$u_t = \rho\ u_{t-1} + \varepsilon_t ,$$

où $|\rho| < 1$, les u_i sont des « termes de perturbation », et les ε_i sont des termes d'erreur qui se conforment aux hypothèses suivantes :

$$E(\varepsilon_t) = 0$$

$$E(\varepsilon_t\ \varepsilon_{t+s}) = \begin{cases} \sigma_\varepsilon^2 & s = 0 \\ 0 & s \neq 0 \end{cases} \qquad \begin{array}{l} \text{pour tout } t \geqslant 0 \\ t, s \text{ nombres} \\ \text{entiers.} \end{array}$$

D'après ce modèle, un *coefficient d'autocorrélation d'ordre s* peut être construit; il est décrit par l'expression

$$\frac{E(u_t u_{t-s})}{\sigma_u^2} = \rho^s. \text{ (Johnston, 1972, p. 245).}$$

Cependant, ce modèle ne peut être appliqué que si les relations entre les termes de perturbation peuvent s'exprimer sous forme d'un polynôme composé des puissances d'une constante ρ :

$$u_t = \rho\ u_{t-1} + \rho^2 u_{t-2} + \ldots + \rho^s u_{t-s} + \varepsilon_t$$

s étant l'ordre jusqu'où il faut mener l'analyse pour que les ε_t soient conformes aux hypothèses déjà énoncées.

Cet ordre est égal à un, plus le nombre de termes intervenant entre u_t et u_{t-s} ; on peut donc le considérer comme mesurant l' « étendue » de l'effet d'autocorrélation.

Cette situation ne se rencontre que rarement dans la réalité, où les données successives manifestent souvent des relations qui ne donnent pas l'impression d'être produites au hasard, mais qui exigent pour leur description des modèles moins restreints que le précédent. Sans un modèle d'autocorrélation spécifié d'avance, il devient difficile de vérifier l'existence ou de mesurer l'autocorrélation dans une série de valeurs résiduelles. Souvent un coefficient d'autocorrélation finit par n'être qu'une sorte de coefficient « moyen » représentant approximativement une relation variable entre les u_t.

Bien qu'un grand nombre des techniques d'analyse d'autocorrélation aient été élaborées dans le contexte de l'analyse de régression, elles peuvent tout aussi bien s'appliquer à n'importe quelle série de données quantitatives. Par exemple, l'indice de contiguïté de Geary (1954, 115-141) fut proposé comme complément à l'analyse de régression de données distribuées dans l'espace, pour signaler l'existence de tendances dans les erreurs résiduelles qui ne peuvent être attribuées au hasard, et pour aider à

décider s'il faut augmenter ou non le nombre de variables indépendantes. Cependant, ce même indice peut tout aussi bien mesurer l'importance des effets de contiguïté dans tout ensemble de données d'une ou de plusieurs variables distribuées dans un groupe de régions mutuellement exclusives, chaque variable ayant une valeur assignée à chaque région.

L'autocorrélation spatiale se distingue de l'autocorrélation en série par le fait que chaque valeur u_t (soit une erreur résiduelle d'une régression, soit tout simplement une valeur assignée à la région t) doit être considérée non seulement en relation avec une valeur u_{t-1} qui la précède ou lui est voisine, mais avec tout un ensemble de valeurs u_i appartenant aux régions qui ont des frontières ou au moins des points en commun avec la frontière de la région t. L'ordre s du coefficient d'autocorrélation a toujours son sens, mais il doit être redéfini en termes du nombre d'unités d'observation qui interviennent sur le plan entre deux régions, au lieu d'être une simple différence de position. On peut souligner la distinction conceptuelle entre les deux types d'analyse en notant qu'une succession de nombres qui manifeste une corrélation en série élevée peut être assignée aux cellules d'un « carrelage » en deux dimensions de façon à produire des configurations de valeurs qui varient énormément dans leur degré d'autocorrélation spatiale.

Par analogie avec le modèle unidimensionnel, le cas de deux dimensions pourrait s'exprimer mathématiquement de cette façon:

$$u_t = \rho \sum_{i=1}^{k_t} u_i + \varepsilon_t ,$$

où k_t est le nombre de régions contiguës à la région t, mais l'hypothèse d'une constante ρ devient irréaliste à cause de l'interaction qui joue dans les deux sens entre les régions. On peut exprimer la relation entre la région t ayant la valeur u_t et une région voisine spécifique t' ayant la valeur $u_{t'}$ de la façon suivante :

$$u_t = \rho\, u_{t'} + \rho \sum_{\substack{i=1 \\ i \neq t'}}^{k_t} u_i + \varepsilon_t .$$

Dans ce cas

$$u_{t'} = \frac{1}{\rho}\, u_t - \sum_{\substack{i=1 \\ i \neq t'}}^{k_t} u_i - \frac{1}{\rho}\, \varepsilon_t$$

et évidemment l'hypothèse d'une constante ρ qui exprime la relation entre toutes paires de régions contiguës ne peut être retenue que dans le cas exceptionnel d'une corrélation parfaite ($\rho = 1$). Toute analyse basée sur le modèle de régression doit donc, au moins pour le moment, rester approximative.

Le problème de l'autocorrélation spatiale a été abordé de deux façons. L'une consiste à employer des modèles basés sur le principe de l'analyse de variance, qui répartissent la variance observée des u_i entre celle attribuable aux relations de contiguïté et celle due aux effets aléatoires. D'autre part, on se sert des modèles qui consistent à réduire un système multidimensionnel en un système unidimensionnel où l'on peut appliquer divers types conventionnels d'analyse de corrélation.

Il semble que la première définition du problème en termes d'analyse de variance soit la proportion von Neumann (1941, 367-395)

$$\frac{\delta^2}{S^2} = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2 / (n-1)}{\sum_{t=1}^{n} (e_t - \bar{e})^2 / n}$$

qui est la différence moyenne au carré des termes d'erreur successifs divisée par la variance d'erreur globale. Si les e_i sont fortement autocorrelés, la différence moyenne au carré compte pour peu dans la variance globale et $\frac{\delta^2}{S^2}$ s'approche de zéro; au fur et à mesure que le degré d'autocorrélation diminue, $\frac{\delta^2}{S^2}$ s'approche de l'unité. Quand n devient grand, la proportion von Neumann adopte approximativement la distribution normale et on peut donc appliquer un test de signification à l'hypothèse d'une corrélation égale à zéro. Cette ligne de pensée a été prolongée avec l'indice de contiguïté de Geary (1954, 115-141) :

$$c = \frac{(n-1)}{2K_1} \cdot \frac{\sum'_{i \neq t} (z_t - z_{t'})^2}{\sum_t (z_t - \bar{z})^2}$$

où k_t = nombre de régions contiguës à la région t

k_1 = somme des k_t

Σ = sommation s'étendant sur toutes les régions

Σ' = sommation s'étendant sur toutes les régions contiguës les unes aux autres

z_i = valeurs standardisées des u_i .

Cet indice a été créé spécifiquement pour des données ayant une distribution spatiale. Pour l'application des tests de signification, on peut employer la formule :

$$R = \frac{1 - c}{M_2}$$

où M_2 = deuxième moment de c autour de l'origine

$$= [n^2k_1^2 + 2n(k_1 + k_2)] \quad [(n-1) / n^2(n-1)\, k_1^2]$$

$$k_1 = \Sigma\, k_t / n , \qquad k_2 = \Sigma\, k_t^2 / n .$$

(King, 1969, 111).

Des indices semblables ont été proposés par Moran (1950, 17-23) et Cliff et Ord (1968). Ces proportions sont relativement faciles à calculer et leur signification peut être déterminée; mais leur interprétation est moins facile parce qu'elles ne sont pas précisément des quotients de deux variances mais plutôt des quotients d'une différence moyenne au carré (pour exprimer la variation à l'intérieur des groupes) et d'une variance globale. En plus, elles ne donnent aucune indication de la répartition dans l'espace des effets d'autocorrélation, n'étant que des proportions sommaires, dont la valeur peut facilement fluctuer avec des changements dans la grandeur ou la forme des régions.

L'autre manière d'aborder le problème consiste à calculer un coefficient d'autocorrélation devant être analogue aux autres coefficients de corrélation déjà existants et avoir certaines qualités statistiques souhaitables. Un tel coefficient peut être élaboré spécifiquement pour les systèmes en deux dimensions. Autrement, l'information de contiguïté qui est contenue dans le système peut être transformée afin de se servir des coefficients paramétriques ou non-paramétriques tels que définis par la statistique ordinaire. Dacey (1969, 479-490) a proposé la formule :

$$r = \frac{M}{A}\left[\frac{\sum_{i=1}^{M}\sum_{j=i+1}^{M} c_{ij}x_i x_j}{x_i^2}\right] = \frac{M}{A} \cdot R$$

où les x_i sont des variables normalisées et standardisées

M est le nombre de régions

$A = 0{,}5 \times \sum_{i=1}^{M}$ (le nombre de relations de contiguïté qui existent avec la région i).

Si M est grand, r s'approche du coefficient de corrélation entre les valeurs associées aux régions contiguës. On peut faire des tests pour vérifier s'il y a des effets non-aléatoires en se servant de la variable standard normale

$$\frac{R - E(R)}{\sqrt{Var(R)}}$$

Cependant, pour le coefficient de Dacey, comme pour les proportions du type « analyse de variance », aucun test n'a été proposé pour déceler la signification de la différence entre deux valeurs distinctes prises par la statistique, i.e. pour examiner l'hypothèse H_0 : $\rho_1 = \rho_2$, $\rho_1 \neq 0$, $\rho_2 \neq 0$, tandis que pour le coefficient de corrélation de Pearson, un tel test existe. Pour cette raison, dans l'essai de simulation et d'analyse des données de mortalité qui sont discutés plus loin, on a décidé d'extraire l'information de contiguïté qui est contenue dans l'arrangement des valeurs en deux dimensions et de la mettre sous une forme convenant à l'analyse

de corrélation conventionnelle, méthode illustrée par le travail d'Olson (1975, 189-204).

La méthode d'Olson consiste essentiellement à exprimer les relations de contiguïté apparaissant dans une carte sous forme de deux séries de données. Les éléments de la première série sont les valeurs associées avec la région étudiée (i.e. dont on examine les relations avec ses régions voisines) et les éléments de la deuxième série sont les valeurs associées avec ces régions voisines. Par exemple, si la région a_i, ayant la valeur v_i possède un côté commun avec les régions a_j, a_k, a_l, ayant des valeurs v_j, v_k, v_i, les éléments suivants apparaîtraient dans les deux séries :

Série 1	*Série 2*
v_i	v_j
v_i	v_k
v_i	v_l
•	•
•	•
•	•

Si on continue ainsi pour toutes les régions a_i, on finit par avoir deux séries de valeurs qui sont analogues aux deux variables X et Y de l'analyse de corrélation ordinaire. Olson voulait examiner la complexité visuelle de cartes comprenant des régions pouvant comporter de une à cinq valeurs (dénommées 1 à 5). Elle construisit une matrice qui résumait les relations entre les régions voisines et calcula un coefficient non-paramétrique de corrélation (le « tau » de Kendall) en se servant des éléments de cette matrice. Mais ses données étaient des nombres entiers; dans la présente étude, les étapes plus avancées de la simulation et de l'analyse demandent des données rationnelles. Des coefficients de corrélation de Pearson (r) ont donc été calculés directement à partir des deux séries.

ESSAI DE SIMULATION

Avant l'analyse des données réelles, un essai de simulation a été fait pour éprouver la méthodologie. Cet essai fournit en outre l'occasion d'observer les correspondances entre certaines configurations de valeurs et les coefficients qu'elles produisent. Les effets d'agglomération et de pondération furent aussi examinés. Aux fins de cette étude, on a disposé des valeurs sur un carrelage 16 x 16 et défini comme contigus les carreaux possédant un côté commun (c'est-à-dire que les « voisins diagonaux » qui n'ont qu'un point commun ont été laissés de côté). Cette définition sera

discutée plus amplement dans le contexte de l'analyse des données de mortalité.

Par analogie visuelle avec les « cartes colorées à tiers égaux » d'Olson, on a construit cinq groupes de dix dispositions chacun, de sorte que ces groupes aient différents degrés d'autocorrélation positive (l'autocorrélation négative a été laissée de côté parce qu'une telle corrélation représente une disposition « en damier », où alternent les valeurs basses et les valeurs hautes, ce qui est rare en géographie). Pour assigner des valeurs aux cellules du carrelage, on a réparti les nombres entiers de 1 à 256 en trois catégories approximativement égales, les plaçant ensuite au hasard dans les cellules. Ce procédé a généré un ensemble de carrelages numérotés 16 x 16 où chaque cellule contenait un nombre unique entre 1 et 256 inclusivement. Deux dispositions ont été préparées pour chaque degré d'autocorrélation positive voulu (nul, faible, faible-moyen, moyen, fort), puis chaque disposition a été permutée pour en obtenir quatre autres, ce qui a donné en tout dix dispositions pour chaque degré de corrélation. La moyenne et la variance de r pour chaque catégorie (prises sur le carrelage 16 x 16) apparaissent dans le tableau 1. Ensuite, un *générateur* de nombres aléatoires sur ordinateur a servi à tirer des échantillons de 256 nombres aléatoires de la distribution standard normale. Chaque échantillon a été trié et ses valeurs numérotées en ordre ascendant, donnant ainsi une liste de rangs correspondant à la liste triée des valeurs ; ces rangs allant de 1 à 256 ont servi à assigner les valeurs aux cellules déjà numérotées. Le résultat final était un ensemble de matrices 16 x 16, chacune contenant un échantillon de valeurs avec une distribution standard normale approximative, et chacune ayant un arrangement de valeurs hautes et de valeurs basses donnant une valeur différente de r.

Le premier essai consistait à réunir les cellules des carrelages 16 x 16 dans des carrelages 8 x 8, 4 x 4, et 2 x 2 successivement, pour examiner comment r varie aux différents niveaux d'agrégation. Chaque cellule agrégée porte la valeur moyenne des cellules qui la composent. On peut supposer que l'ensemble des valeurs moyennes a aussi une distribution standard normale approximative, car ses éléments ne sont que des combinaisons linéaires des valeurs originales. À chaque niveau d'agrégation, un coefficient de corrélation a été calculé pour chaque nouvelle disposition ; les moyennes et les variances pour chaque niveau d'agrégation apparaissent dans le tableau 1. Évidemment les valeurs de r vont en diminuant au fur et à mesure qu'on passe des carrelages 16 x 16 aux 8 x 8 et aux 4 x 4. Ceci s'accorde bien avec l'intuition, car un grand nombre des relations de ressemblance entre les cellules qui contribuèrent à augmenter la valeur de r pour les carrelages 16 x 16 sont absorbées dans les cellules plus grandes. L'augmentation dans les valeurs moyennes de r au niveau 2 x 2 présente une énigme, mais elle n'a aucune signification au seuil de confiance de 95 pour cent et peut donc être considérée comme un artifice mathématique créé par la perte de degrés de liberté dans les calculs.

Tableau 1

Moyenne et variance de R, cellules carrées et égales

Disposition	Moyenne	Variance
r fortement positif		
16 x 16	0,8726	0,0044
8 x 8	0,8146	0,0069
4 x 4	0,5673	0,0320
2 x 2	0,6675	0,0104
r moyennement positif		
16 x 16	0,6467	0,0017
8 x 8	0,5454	0,0037
4 x 4	0,2204	0,0216
2 x 2	0,4212	0,0032
r faible à moyennement positif		
16 x 16	0,5626	0,0030
8 x 8	0,4198	0,0043
4 x 4	0,0799	0,0156
2 x 2	0,3400	0,0042
r faiblement positif		
16 x 16	0,3508	0,0031
8 x 8	0,2242	0,0261
4 x 4	−0,0397	0,0189
2 x 2	0,1656	0,0183
r presque nul (aucune corrélation)		
16 x 16	0,0460	0,0020
8 x 8	−0,1479	0,0094
4 x 4		
2 x 2	−0,1507	0,0075

Les tests de signification ont été faits entre les coefficients r_1 et r_2 pour divers degrés de corrélation et pour toutes les paires de niveaux d'agrégation, en se servant de la formule suivante (King, 1969, 131-132) :

$$u = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \quad \text{pour examiner } H_0 : \rho_1 = \rho_2$$

où ρ_1 , ρ_2 sont les coefficients de corrélation de la population,

n_1 , n_2 sont les nombres de cellules sur lesquels sont basés r_1 et r_2 ,

z_1 , z_2 sont les valeurs prises par le " z " de Fisher,

$$z = \frac{1}{2} \ln \frac{1 + r}{1 - r}$$

qui a la distribution standard normale. La statistique u a approximativement la distribution standard normale. Les résultats pour les comparaisons des coefficients 16 x 16 et 8 x 8 se sont montrés peu réguliers, mais en passant du niveau 16 x 16 au niveau 4 x 4, presque tous les coefficients ont manifesté un changement significatif au seuil de signification de 0,05. Le manque de signification dans l'augmentation observée entre les niveaux 4 x 4 et 2 x 2, en plus de la croissance de la variance qui se produit au niveau 4 x 4, mettent en question le nombre minimum de cellules requis pour un calcul fiable d'une mesure d'autocorrélation spatiale. Dans le contexte de la présente étude, ce problème ne peut être résolu, mais il a néanmoins une importance critique dans l'analyse de données empiriques qui ne sont souvent disponibles que pour un petit nombre de régions.

Le deuxième essai consistait à combiner les 256 cellules originales en 50 régions rectangulaires de grandeur et de forme diverses (figure 1), pour examiner les effets de pondération des relations de contiguïté sur l'importance numérique de r. Deux règles de pondération ont été essayées : la pondération proportionnelle à la longueur de la frontière partagée, et la pondération proportionnelle à la racine carrée du quotient de la superficie des deux régions en question (la pondération proportionnelle au quotient des superficies, sans modification, a donné des autocorrélations négatives pour presque toutes les dispositions, résultat qui n'avait aucun sens). En plus, on a essayé une pondération combinée, le produit des deux facteurs obtenus des pondérations précédentes, ainsi qu'un calcul non pondéré pour fins de comparaison. Les moyennes et les variances des valeurs de r ainsi obtenues apparaissent au tableau 2.

Figure 1

RÉGIONS RECTANGULAIRES
ESSAI DE SIMULATION

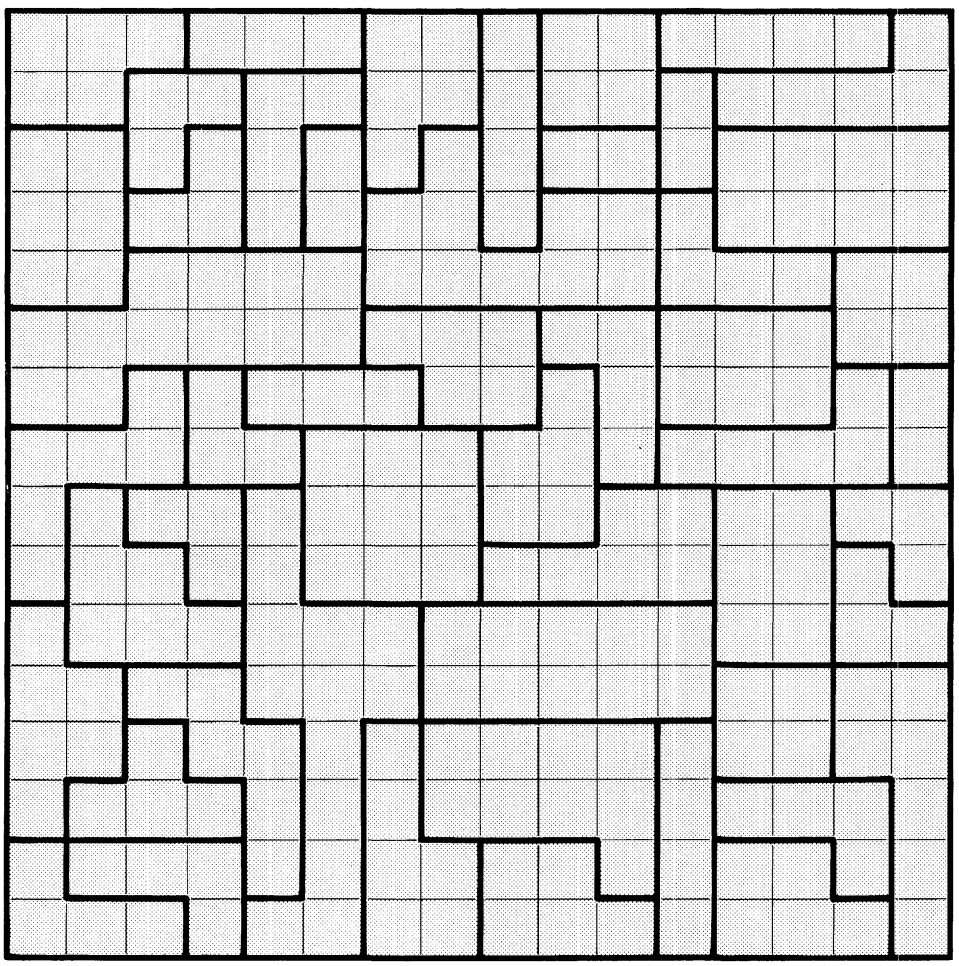


Tableau 2

Moyenne et variance de R,

Régions rectangulaires avec pondération

Disposition	Moyenne	Variance
r fortement positif		
Sans pondération	0,7479	0,0131
Pondération 1	0,8362	0,0065
Pondération 2	0,6120	0,0283
Pondération combinée	0,9261	0,0017
r moyennement positif		
Sans pondération	0,4458	0,0080
Pondération 1	0,6110	0,0059
Pondération 2	0,2420	0,0106
Pondération combinée	0,8275	0,0013
r faible à moyennement positif		
Sans pondération	0,3045	0,0096
Pondération 1	0,4985	0,0082
Pondération 2	0,0932	0,0132
Pondération combinée	0,7716	0,0023
r faiblement positif		
Sans pondération	0,1537	0,0248
Pondération 1	0,3742	0,0241
Pondération 2	−0,0547	0,0220
Pondération combinée	0,7040	0,0067
r presque nul (aucune corrélation)		
Sans pondération	−0,0754	0,0059
Pondération 1	0,1962	0,0143
Pondération 2	−0,2462	0,0102
Pondération combinée	0,5999	0,0042

NOTE : Pondération 1 = pondération par longueur de frontière partagée.
Pondération 2 = pondération par la racine carrée du quotient des superficies des deux régions.
Pondération combinée = produit de 1 et 2.

Le procédé de pondération consiste à augmenter ou à diminuer la différence absolue entre chaque paire de valeurs X et Y en la multipliant par un facteur de pondération appropriée. En effet, tout un ensemble de paires X–Y a été créé, où chaque nouvelle paire a conservé la même moyenne $\frac{X + Y}{2}$ que la paire originale, mais les quantités de la nouvelle paire ont été rapprochées ou éloignées pour en augmenter ou en diminuer l'effet de contiguïté, selon les exigences du schéma de pondération. Un examen de la formule utilisée pour calculer r :

$$r = \frac{\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[\Sigma X^2 - (\Sigma X)^2]\ [\Sigma Y^2 - (\Sigma Y)^2]}}$$

confirme que ce procédé accomplit la pondération de r proportionnellement à chacun des rapports de contiguïté.

On voit facilement dans le tableau 2 que les valeurs de r non pondérées pour les 50 régions rectangulaires sont un peu plus basses que les valeurs pour les carrelages 16 x 16.

On s'attendait à ce résultat d'après la discussion précédente des effets d'agrégation. La pondération par longueur de frontière commune tend à augmenter ces valeurs, mais pas jusqu'au point où elles seraient égales aux valeurs obtenues pour les carrelages 16 x 16 ; cette tendance peut facilement s'expliquer si l'on note que la longueur moyenne de frontière commune (1,97) est plus grande que l'unité. La pondération par la racine carrée du quotient des superficies, au contraire, a tendance à donner des valeurs plus basses, ce phénomène est fort difficile à expliquer. En effet, on s'attendrait à voir très peu de changements entre ces valeurs-ci et les valeurs non pondérées, car chaque r contient un terme pondéré par le facteur $\sqrt{s_i / s_j}$, s_i, s_i étant les superficies des régions i et j respectivement, et aussi un terme pondéré par le facteur $\sqrt{s_j / s_i}$. Pour le moment, on va se borner à remarquer l'effet de cette pondération, sans proposer d'explication. La pondération combinée a donné des valeurs très élevées pour r (avec une moyenne de 0,5999 pour les arrangements supposés sans corrélation). Ceci sert principalement à avertir le lecteur que les pondérations combinées, bien qu'intéressantes comme possibilité théorique, ne peuvent être tentées en pratique sans qu'on connaisse d'abord la nature et le degré de l'interaction entre les facteurs de pondération. Dans le cas actuel, on peut facilement suggérer que la longueur de la frontière commune et le quotient des superficies peuvent bien être positivement corrélés. Par conséquent le facteur de pondération combiné pourrait bien compter deux fois une certaine partie des effets de contiguïté. Des tests de signification faits au moyen de la formule " u " ont montré que presque toutes les différences entre les valeurs de r obtenues par différentes pondérations du même arrangement de valeurs étaient significatives au seuil de 0,05.

ANALYSE DES DONNÉES SUR LA MORTALITÉ DUE AUX MALADIES CHRONIQUES

Pour illustrer l'application de cette méthodologie à un problème du monde réel, on a examiné l'existence possible d'une autocorrélation spatiale, dans les taux de mortalité, pour les maladies chroniques. Les maladies chroniques sont celles qui durent longtemps, souvent sous une forme aiguë. Leur étiologie est complexe et mal connue, mais on se doute que les facteurs du milieu y jouent un rôle important, ce qui donnerait à leur distribution une certaine signification spatiale.

Le département d'Aide sociale de la Ville de Montréal nous a fourni, pour l'année 1972, le nombre de personnes décédées à cause de maladies chroniques, pour chacun des 291 secteurs de recensement de la ville. Pour fins de comparaison, le nombre absolu de décès dans chaque secteur a été converti en un taux de mortalité pour 1000 personnes (afin d'enlever l'effet de la variation dans la population totale des secteurs). Ce taux a été standardisé selon l'âge (afin d'enlever l'effet de la variation dans la répartition parmi les catégories d'âge) [1].

La standardisation par sexe a été jugé non nécessaire après un examen préliminaire des données.

La méthode directe de standardisation dont on s'est servi ici, calcule le taux de mortalité qui serait survenu dans une population de référence (population standard), si les groupes d'âge de cette population étaient exposés au même risque qui existe dans la population à l'étude. Si on a M groupes d'âge, on procède de la façon suivante pour chaque groupe i, $i = 1, \ldots, M$.

$$\frac{N_i \times T_i}{1000} = A_i$$

où N_i = nombre de personnes dans le groupe i de la population standard

T_i = taux de mortalité pour le groupe i de la population étudiée

A_i = nombre de décès à attendre dans le groupe i de la population standard si le taux de mortalité T_i s'y appliquait

Puis $$\frac{\sum_{i=1}^{M} A_i}{P} = M$$

où P = la population standard

M = taux de mortalité standardisé pour la population d'étude, à tous les âges.

---

[1] Pour une discussion plus ample des taux standardisés de mortalité et de la nécessité de la standardisation voir Bradford Hill (1971). *Principles of Medical Statistics,* 9ième édition, pp. 203 ff.

En répétant ces calculs pour chaque région d'observation (secteur de recensement ou groupes des secteurs), on obtient un ensemble de taux de mortalité dont l'influence prépondérante des différences entre les proportions des groupes d'âge a été enlevée.

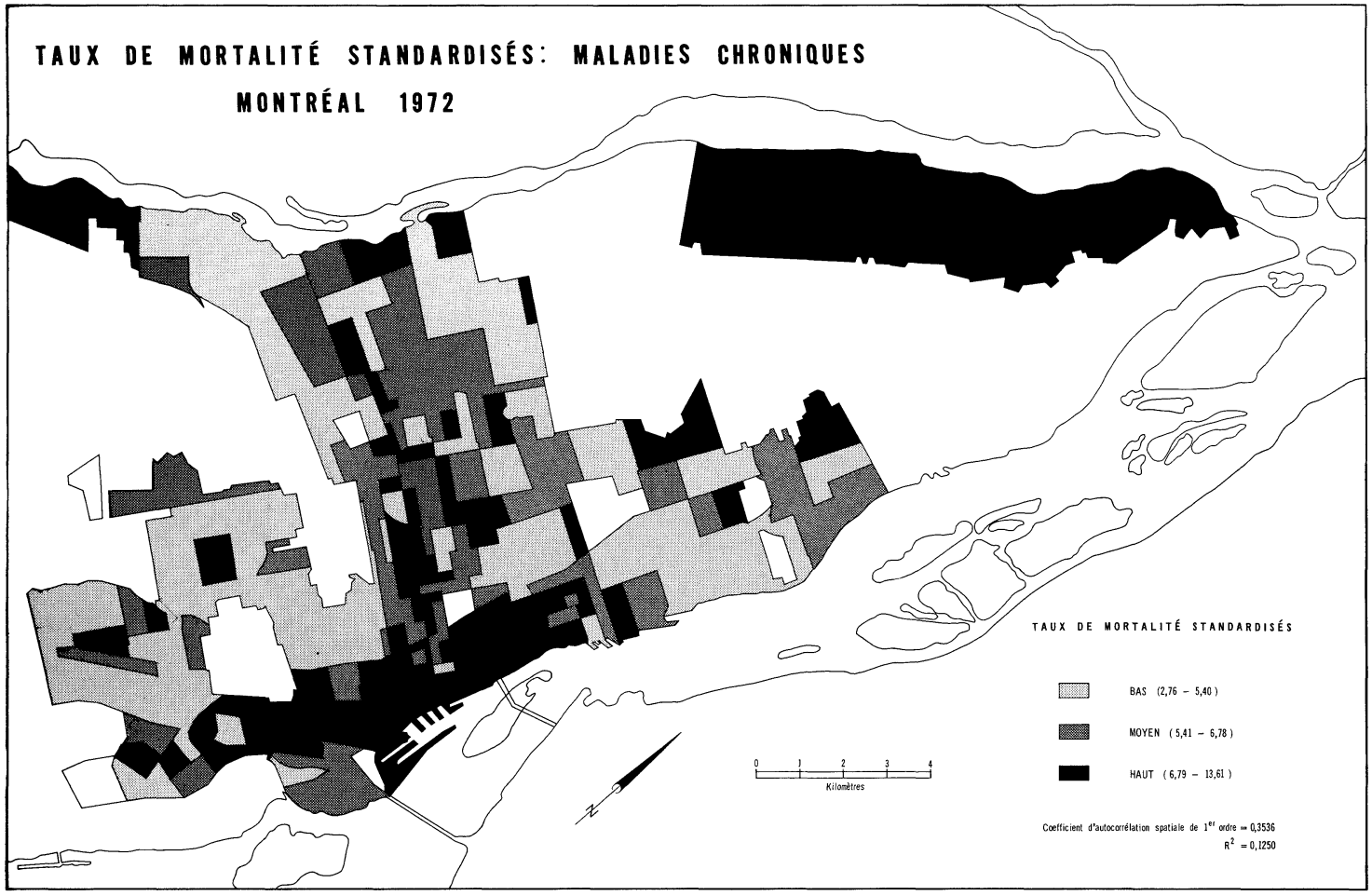
Aux fins de cette étude, la population fut divisée en six groupes d'âge : 0-34 (ans), 35-44, 45-54, 55-64, 65-69, et 70 ans et plus. Cette répartition fait ressortir la tendance des maladies chroniques à frapper surtout les personnes plus âgées ; elle représente aussi la meilleure combinaison de différentes catégories d'âge présentées dans les données municipales de mortalité et dans les données de recensement.

Les biostatisticiens sont généralement d'accord pour dire (Hill, 1971) que le procédé de standardisation dont on se sert ici requiert un minimum approximatif de 20 décès dans chaque région d'observation afin d'éliminer l'influence excessive des incidences isolées sur les calculs, et que les régions en ayant moins de 20 doivent être combinées pour atteindre ce minimum. Cependant, l'essai de simulation montre que le coefficient d'autocorrélation spatiale employé ici peut être « déséquilibré » quand les régions sont combinées, et qu'il faut donc procéder avec prudence. Le choix d'un minimum de 15 décès par secteur a entraîné la combinaison des 291 secteurs de recensement en 241 régions de taux de mortalité standardisés (''région TMS''), ce qui réduit le nombre de régions d'à peu près 15 pour cent. Les résultats de l'essai de simulation indiquent que l'effet d'une telle réduction devrait être minime.

Pour cet ensemble de données, le coefficient d'autocorrélation spatiale du premier ordre, calculé en définissant la contiguïté comme la possession d'une frontière commune, atteint une valeur de 0,3536. Ce coefficient, quoique bas, est néanmoins significatif au seuil de 0,01 grâce au grand nombre de degrés de liberté disponible pour les calculs. Le coefficient de détermination, (R^2) est 0,1250, ce qui indique que 12,5% de la variation dans les taux de mortalité standardisés peut être attribuée aux effets de contiguïté (figure 2). Malgré le faible degré d'explication fourni par le coefficient d'autocorrélation spatiale, plusieurs tendances de concentration se dégagent de la figure 2. Une région de mortalité élevée s'étend du centre-ville au sud-ouest vers le canal Lachine, tandis que les quartiers de l'ouest et du nord-ouest de Montréal démontrent des taux de mortalité plus bas. Cependant, les secteurs au nord et à l'est du centre-ville présentent une mosaïque de valeurs, ce qui explique probablement la faiblesse du coefficient obtenu. Le manque de données pour les municipalités voisines n'a pas permis de faire ressortir suffisamment les tendances spatiales dans ces secteurs.

Comme le facteur spatial n'explique qu'une petite partie de la variation dans la mortalité due aux maladies chroniques, on est tenté de chercher d'autres facteurs étiologiques, probablement d'un caractère non spatial. Quoique les processus sous-jacents qui entraînent les maladies chroniques

Figure 2


TAUX DE MORTALITÉ STANDARDISÉS: MALADIES CHRONIQUES
MONTRÉAL 1972
TAUX DE MORTALITÉ STANDARDISÉS
BAS (2,76 – 5,40)
MOYEN (5,41 – 6,78)
HAUT (6,79 – 13,61)
0 1 2 3 4
Kilomètres
N
Coefficient d'autocorrélation spatiale de 1er ordre = 0,3536
R2 = 0,1250

soient encore mal connus, des articles récents (Girt, 1973 ; Cairns, 1975, 64-75) suggèrent que la susceptibilité individuelle et l'exposition à long terme aux facteurs étiologiques du milieu sont probablement responsables. Comme le dit Girt (1973) dans son étude sur la bronchite chronique, le milieu où demeure actuellement une personne est probablement moins significatif que ses contacts antécédents à des milieux nocifs comme facteur dans l'étiologie de la bronchite chronique et, que les effets de cette histoire personnelle peuvent agir à très long terme. Nous pouvons donc conclure que la majeure partie de la variation dans la mortalité due aux maladies chroniques, inexpliquée par la contiguïté spatiale, est probablement due aux variations individuelles de susceptibilité, d'hérédité, et d'histoire médicale et personnelle. La quantité d'autocorrélation spatiale qui s'y trouve, petite mais néanmoins fortement significative, représente donc l'influence des conditions du milieu actuel, et probablement le fait que les personnes ayant des antécédents semblables d'exposition aux pathogènes dans le milieu tendent à se réunir dans les mêmes quartiers sous l'influence des facteurs socio-économiques.

Si l'on voulait pousser plus loin l'analyse de cet ensemble de données, on pourrait effectuer le calcul des coefficients d'autocorrélation spatiale de deuxième ordre, de troisième ordre, et ainsi de suite. On pourrait aussi expérimenter avec des pondérations basées sur la superficie des régions ou la longueur des frontières, ou avec une définition « diagonale » de contiguïté (le fait d'avoir une frontière ou même un seul point en commun). Le choix de la pondération et de la définition de la contiguïté devrait, dans le cas idéal, avoir sa justification dans la nature du phénomène étudié. La pondération par longueur de frontière partagée convient à un phénomène qui exerce son influence en « s'écoulant » à travers des frontières, tandis que la pondération par superficie convient à un phénomène qui influence des régions voisines par son importance et sa proximité même, sans égard aux frontières spécifiques. Une discussion parallèle nous montre qu'une définition latérale de contiguïté (le fait d'avoir une frontière en commun) convient aux phénomènes « d'écoulement » et une définition diagonale aux phénomènes qui agissent par importance et proximité. Malheureusement, la théorie de l'étiologie des maladies chroniques n'est pas encore assez bien établie pour nous permettre de justifier un choix de pondération ou une définition de contiguïté de préférence à un autre, mais une analyse comparative pourrait donner des résultats intéressants.

CONCLUSION

Cette analyse, bien que préliminaire, illustre l'utilité de l'analyse d'autocorrélation spatiale pour obtenir une compréhension plus approfondie des problèmes dans le domaine de la santé, surtout quand les méthodes quantitatives peuvent compléter la théorie médicale. Au cours de cette étude, quelques problèmes ont aussi été mis à jour, problèmes qu'il faut résoudre avant de multiplier des analyses de ce genre. Beaucoup de difficul-

tés seraient évitées si les données dans le domaine de la santé étaient facilement disponibles, plus exhaustives et de bonne qualité. Par exemple, il se peut que des données sur un phénomène ne soient recueillies que pour un petit nombre de régions vastes et hétérogènes, ce qui donne des résultats peu fiables (à cause de la perte des degrés de liberté) et difficiles à interpréter; ou bien, les régions peuvent être plus petites mais contenir peu d'observations du phénomène, et donc des petites variations numériques peuvent exercer une influence hors de proportion sur les résultats. Du côté méthodologique, le coefficient d'autocorrélation spatiale, bien qu'il ait spécifiquement incorporé de l'information spatiale, n'est qu'un chiffre isolé. Toute indication de la répartition spatiale des valeurs en est disparue. Il semble que la cartographie et l'interprétation subjective de la répartition des données soient toujours des compléments essentiels à l'analyse d'autocorrélation (voir la figure 2). Les problèmes d'agrégation spatiale et de pondération ont fait l'objet d'une étude assez poussée au cours de l'essai de simulation, mais ils exigent encore plus de recherches avant que l'analyse de l'autocorrélation spatiale ne devienne un outil dont on puisse se servir en toute sûreté.

BIBLIOGRAPHIE

CAIRNS, J. (1975) The Cancer Problem. *Scientific American*, 233 (5) : 64-75.

CLIFF, A.B., and ORD, J.K., (1968) *The Problem of Spatial Autocorrelation*. Université de Bristol, Département de géographie.

DACEY, M.F. (1969) A Review on Measures of Contiguity for Two and K-Color Maps. In Berry, B.J.L., and Marble, D.F. (1969) *Spatial Analysis : A Reader in Statistical Geography*. Englewood Cliffs, Prentice-Hall, 479-490.

DRAPER, N.R., and SMITH, H. (1966) *Applied Regression Analysis*, New York, Wiley, 407 p.

DURBIN, J. and WATSON, G.S. (1950 et 1951) Testing for Serial Correlation in Least-Squares Regression. *Biometrika*, 37 : 409-428 et 38 : 159-178.

EZEKIEL, M. and FOX, K.A. (1959) *Methods of Correlation and Regression Analysis*, 3ième édition, New York, Wiley, 548 p.

FREUND, J. (1971) *Mathematical Statistics* 2ième édition, Englewood Cliffs, Prentice-Hall, 464 p.

GEARY, R.C. (1954) The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, 5 : 115-141 ; réimprimé in Berry, B.J.L., and Marble D.F. (1969) *Spatial Analysis : A Reader in Statistical Geography*. Englowood Cliffs, Prentice-Hall, 461-478.

GIRT, J.L. (1973) Simple Chronic Bronchitis and Urban Ecological Structure. In McGlashan, N.D. (1973) *Medical Geography : Techniques and Field Studies*, London. Methuen, 336 p.

HILL, A.B. (1971) *Principles of Medical Statistics*. 9ième édition. New York, Oxford University Press, 390 p.

JOHNSTON, J. (1972) *Econometric Methods*. 2ième édition, New York. McGraw-Hill, 437 p.

KING, L.J. (1969) *Statistical Analysis in Geography*. Englewood Cliffs, Prentice-Hall, 288 p.

MORAN, P.A. (1950) Notes on Continuous Stochastic Phenomena. *Biometrika*, 37 : 17-23.

OLSON, J.M. (1975) Autocorrelation and Visual Map Complexity. *Annals of the Association of American Geographers* 65 (2) : 189-204.

VON NEUMANN, J. (1941) Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Ann. Math. Statist.* 12 : 367-395.

RÉSUMÉ

BOUCHARD, Diana C. : L'autocorrélation spatiale et les données de santé ; une étude préliminaire.

L'analyse de l'autocorrélation spatiale cherche à mesurer jusqu'à quel point la variation dans un ensemble de données réparties dans l'espace est due aux relations de

contiguïté. Du point de vue mathématique, il existe deux façons d'aborder le problème : l'analyse de variance et le calcul d'un coefficient d'autocorrélation. Dans cette étude, une méthode du deuxième type est appliquée d'abord à un ensemble de carrelages d'essai possédant divers degrés d'autocorrélation spatiale et puis à la distribution spatiale de mortalité due aux maladies chroniques, à Montréal, en 1972. On conclut qu'elles révèlent une autocorrélation faible mais significative par rapport aux données de mortalité, et que d'autres facteurs suggérés dans la littérature récente de la géographie médicale pourraient bien avoir plus d'influence que la contiguïté spatiale elle-même.

MOTS-CLÉS : Méthodes quantitatives, autocorrélation, contiguïté spatiale, géographie médicale, mortalité, maladies chroniques. Ville de Montréal.

ABSTRACT

BOUCHARD, Diana C. : Spatial Autocorrelation and Health Care Data : A Preliminary Study.

Spatial autocorrelation analysis attempts to measure the extent to which variation in spatially distributed data is due to the existence of contiguity relationships. From a mathematical point of view there are two general approaches to the problem : analysis of variance, and the calculation of a coefficient of spatial autocorrelation. In this study a method of the second type is applied, first to a series of test patterns with varying degrees of spatial autocorrelation, and then to the spatial distribution of chronic disease mortality in Montreal in 1972. The conclusion of the mortality data analysis were that a slight but significant autocorrelation effect was present and that other factors indicated in the recent medical geography literature could well be more influental than spatial contiguity itself.

KEY WORDS : Quantitative Methods, Autocorrelation, Contiguity, Weighting, Medical Geography, Chronic Diseases, Mortality. City of Montreal.