International Journal of Geo-Information

MDPI

Article

Amélioration des indices d'autocorrélation spatiale, des méthodes d'estimation et de modélisation spatiale par standardisation sur la distance

Marc Souris [1,2,*] et Florent Demoraes [3]

[1] UMR Unité des Virus Emergents (UVE : Aix-Marseille Univ—IRD 190—Inserm 1207—IHU Méditerranée Infection), 13005 Marseille, France

[2] RS&GIS FoS, School of Engineering and Technology, Asian Institute of Technology, P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand[3] Univ Rennes, CNRS, ESO—UMR 6590, F-35000 Rennes, France ; florent.demoraes@univ-rennes2.fr

* Correspondance : marc.souris@ird.fr

Reçu : 5 mars 2019 ; Acceptée : 22 avril 2019 ; Publié : 24 avril 2019 – version originale en français

Résumé : Etant donné un ensemble de points dans un espace de dimension supérieure à 1, la distribution statistique du nombre de couples de points en fonction de leur distance n'est pas constante. Cette distribution n'est pas prise en compte dans un grand nombre de méthodes classiques utilisées en analyse spatiale et basées sur des moyennes, comme les indices d'autocorrélation spatiale, les méthodes d'interpolation par noyau ou les méthodes de modélisation spatiale (autorégressive ou géographiquement pondérée). Cette distribution a un impact direct sur les calculs et les résultats des indices et des estimations et en ne tenant pas compte de cette distribution des distances, les calculs d'analyse spatiale peuvent être biaisés. Dans cet article, nous introduisons une "standardisation spatiale", qui corrige et ajuste les calculs par rapport à la distribution des couples de points en fonction de leurs distances. A titre d'exemple, nous appliquons cette correction au calcul des indices d'autocorrélation spatiale (indices de Moran et de Geary) et au calcul de surface de tendance (par interpolation spatiale par noyau) sur les résultats de l'élection présidentielle française de 2017.

Mots clés : analyse spatiale ; autocorrélation spatiale ; modélisation spatialisée ; interpolation spatiale par noyau ; standardisation ; SD-correction.

1. Introduction

De très nombreux phénomènes naturels ou anthropiques présentent ce que l'on appelle « une dépendance spatiale » : les différentes valeurs d'une variable localisée liée à ce phénomène ne sont pas indépendantes entre elles, ce qui signifie que deux valeurs proches ont plus tendance à se ressembler que deux valeurs éloignées. On constate ainsi dans de très nombreux phénomènes une augmentation de la variance en fonction de la distance. Cette dépendance spatiale est d'ailleurs considérée comme le fondement de la géographie (« everything is related to everything else, but near things are more related than distant things » [TOB 70]). Cette dépendance spatiale concerne de nombreux domaines scientifiques, comme par exemple la géologie, la botanique, l'économie, l'épidémiologie, la météorologie...

La dépendance spatiale entre valeurs numériques d'une variable localisée peut s'exprimer par le concept d'autocorrélation spatiale, qui permet de formuler la corrélation entre les valeurs X_i des objets P_i en fonction des relations métriques ou topologiques entre les objets :

« Etant donné un ensemble de n unités géographiques, on appelle autocorrélation spatiale la relation constatée, pour les $n(n-1)/2$ paires d'unités, entre les différences des valeurs d'une variable mesurée en ces lieux et une mesure de la proximité géographique » [2-3].

Pour estimer cette autocorrélation spatiale, on utilise des indices numériques, à l'image d'un indice de corrélation classique en dimension 1 : un indice d'autocorrélation spatiale est une mesure statistique des corrélations entre valeurs d'objets localisés utilisant les relations métriques ou topologiques entre ces objets.

Tester la significativité statistique de l'autocorrélation spatiale est le plus efficace pour montrer l'existence d'une dépendance spatiale. Les indices d'autocorrélation permettent ainsi d'étudier l'agrégation ou la dispersion spatiale globale ou locale des valeurs et de mesurer le rôle de la contiguïté ou de la distance dans les interactions spatiales. Un indice d'autocorrélation spatiale peut être une mesure moyenne globale ou ne concerner que le voisinage d'un lieu [4–7].

De nombreux indices d'autocorrélation spatiale ont été développés depuis 70 ans [2-3,8]. Ces indices sont construits à partir de la relation géométrique ou topologique entre les couples de points (P_i, P_j) d'une part et de la différence des valeurs X_i, X_j des deux points du couple d'autre part. La relation géométrique ou topologique entre les deux points d'un couple doit exprimer la dépendance spatiale et être transformée en une valeur numérique afin de construire un indice numérique. Pour cela, plusieurs options sont possibles, en fonction de la manière dont on souhaite prendre en compte la dépendance spatiale [9-12] :

- Par la prise en compte des relations de voisinage. On peut utiliser le voisinage direct (au sens de Voronoï) entre les deux points - ou centroïdes dans le cas de polygones - du couple en affectant 1 si les deux points sont voisins, 0 sinon. Lorsque l'on veut prendre en compte les relations de contiguïté ou d'adjacence, on peut utiliser la longueur du bord commun entre les deux objets (soit de la tessellation de Voronoï dans le cas d'un semis de points, soit la longueur de la frontière dans le cas de zones adjacentes).
- Par la prise en compte de la distance entre les objets (représentés par des points ou centroïdes P_i). On utilise alors une fonction de la distance (distance euclidienne, distance de Manhattan, distance le long d'un réseau valué, etc.). Elle est souvent limitée à une distance maximale dite « d'influence » (notée dans la suite $dmax$), au-delà de laquelle la valeur vaut 0, signifiant une absence de dépendance spatiale au-delà de cette distance. Cette fonction peut être polynômiale, par exemple $\max\left(0, 1 - \frac{d(P_i,P_j)^k}{d_{max}{}^k}\right) avec\ k = \frac{1}{2}, 1, 2 \ldots$-, gaussienne (par exemple $\exp(-d(P_i,P_j)^2/dmax^2)$), sigmoïde, etc. La distance maximale $dmax$ peut être fixée pour tous les couples ou être dépendante d'un caractère lié à la densité. Par exemple, $dmax$ peut être fonction de la distance au n-plus proche voisin de l'un des points du couple). On peut l'estimer par la portée du semi-variogramme correspondant à la situation à analyser.

Ces valeurs liées aux couples d'objets (P_i, P_j) sont appelées poids spatiaux, notés w_{ij}. Ils sont souvent regroupés dans une matrice W, symétrique positive de diagonale nulle, et symétrique si, $\forall i, j \quad w_{ij} = w_{ji}$ (lorsque les relations spatiales sont symétriques). Les poids spatiaux sont fondamentaux dans les calculs d'autocorrélation spatiale puisqu'ils expriment de façon numérique la dépendance spatiale. Lorsque l'on prend en compte les relations de voisinage, on parle de matrice de contiguïté. Lorsque l'on utilise les distances, on parle de matrice de distance.

Plusieurs indices sont utilisés pour mesurer l'autocorrélation spatiale. La plupart de ces indices sont des moyennes sur l'ensemble des couples d'objets et dérivent de l'indice décrit par Mantel [13]. Ils supposent donc implicitement que le phénomène est stationnaire, c'est-à-dire qu'il ne dépend pas du lieu et qu'il correspond à un processus global.

Le plus utilisé est l'indice de Moran [2-4]. Il est défini comme la moyenne des produits des valeurs normalisées des couples de points pondérés par le poids spatial. L'indice de Moran correspond à un indice de corrélation classique (Pearson) étendu aux objets voisins, et muni de la pondération spatiale W. Il utilise ainsi un modèle multiplicatif :

$$I_{Moran} = \frac{1}{S}\sum_{i,j} w_{ij}(\frac{X_i - m}{\sigma})(\frac{X_j - m}{\sigma}) \quad (1)$$

où m est la moyenne des valeurs X_i de l'ensemble des objets, σ l'écart-type des X_i, w_{ij} le poids spatial du couple de point (P_i, P_j), et S la somme des poids spatiaux ($S = \sum_{i,j} w_{ij}$).

L'espérance de l'indice de Moran sous l'hypothèse nulle (pas d'autocorrélation spatiale) est :

$$E(I_{Moran}) = \frac{-1}{N-1} \quad (2)$$

où N est le nombre de points. La variance dépend de W et est donnée dans [9].

Dans la littérature, l'indice est souvent présenté par la formule équivalente mais moins intelligible :

$$I_{Moran} = \frac{N}{S}\sum_i \sum_j w_{ij}(X_i - m)(X_j - m) / \sum_i (X_i - m)^2 \quad (3)$$

Autre indice d'autocorrélation spatiale largement utilisé, l'indice de Geary [2,3,5] est lui construit sur un modèle additif plutôt que multiplicatif : il est défini comme la moyenne des carrés des différences des valeurs normalisées des couples de points :

$$I_{Geary} = \frac{1}{S}\sum_{i,j} w_{ij}(\frac{X_i - m}{\sigma} - \frac{X_j - m}{\sigma})^2 \quad (4)$$

D'autres indices d'autocorrélation spatiale globale n'utilisent pas de variables normalisées, et sont donc moins utilisés : *Black Black Seal*, *Black White Join*, *Knox* [2].

Enfin, d'autres sont utilisés pour estimer l'autocorrélation locale en un point P_i, et sont connus sous le nom d'indices d'association spatiale locale (LISA) [6]. Par exemple, l'indice de Moran local est donné par la formule :

$$I_{Moran}(P_i) = \frac{1}{S_i}\frac{(X_i - m)}{\sigma}\sum_{j,j\neq i} w_{ij}\left(\frac{X_j - m}{\sigma}\right), \qquad avec\ S_i = \sum_{j,j\neq i} w_{ij} \quad (5)$$

L'indice de Getis–Ord [14] est un autre exemple de LISA. Il est construit comme l'indice de Moran local, mais de façon à en faire un Z-score (c'est-à-dire le nombre d'écart-type qui sépare la valeur de la moyenne attendue) [2–3].

D'autres opérations d'analyse spatiale utilisent une moyenne ou une somme de valeurs pondérées avec des poids qui sont fonction de la distance. Il s'agit notamment de processus d'estimation par noyau et de modélisation statistique prenant en compte l'autocorrélation spatiale, comme les modèles de régression simultanément autorégressif (SAR), les modèles de régression conditionnellement autorégressifs (CAR), ou les modèle de régression avec poids spatial (GWR) [2–3,8,12, 15]) :

- L'estimation par noyau (Kernel estimation et Kernel Density estimation) étend à la dimension 2 les principes de l'estimation par noyau en dimension 1. Lorsque la variable est numérique, l'interpolation par noyau consiste à effectuer en chaque point d'une grille la moyenne des valeurs pondérées par une fonction de la distance (appelé ici *noyau*) pour tous les objets situés à une distance inférieure à une distance donnée $dmax$ [16]. Le noyau est une fonction de la distance d de l'objet au point de la grille et est utilisé pour calculer le poids spatial. On peut prendre une fonction linéaire (par exemple $(dmax - d)/dmax$), une fonction quadratique (par exemple $\left(\frac{dmax-d}{dmax}\right)^2$), une gaussienne (par exemple $\frac{1}{\sqrt{2\pi}} e^{-1/2(\frac{dmax-d}{dmax})^2}$), etc. Lorsque la variable est qualitative, l'estimation des densités par noyau (Kernel Density Estimation) consiste à effectuer en chaque point d'une la grille la somme pondérée des effectifs pour tous les objets situés à une distance inférieure à une distance donnée $dmax$, chaque objet étant pondéré par le noyau.
- Les modèles spatiaux autorégressifs (Autoregressive Regression, Simultaneous Autoregressive Regression, Conditional Autoregressive Regression, Generalized Additive Model, Structured Additive Regression) utilisent également une matrice de poids spatiaux construite comme pour

les indices d'autocorrélation spatiale [2–3][16–18]. Par exemple, pour les régressions autorégressives, on a :

$$z_j = \sum_k x_{jk}\beta_k + \rho \sum_{i,i\neq j} w_{ij} z_i + \varepsilon_j \qquad (Z = X\beta + \rho WZ + \varepsilon)$$

où z_j est la variable dépendante au point P_j, x_{jk} les variables explicatives au point P_j, w_{ij} un poids spatial dépendant de la distance entre les points P_i et P_j, et ρ un paramètre du modèle permettant de modéliser la force et la nature de la dépendante spatiale (attraction ou répulsion).

Parfois les poids spatiaux correspondent à une pondération relative et non absolue entre les voisins : pour tous les individus j, la somme des poids de tous ses voisins $\sum_i w_{ij}$ est alors égale à 1, et $\sum w_{ij} z_i$ est une moyenne pondérée. On dit alors que la matrice W est standardisée sur les lignes. Si tous les poids sont égaux, cela revient à ajouter dans le modèle la moyenne des valeurs voisines. Ils peuvent également avoir une influence absolue : dans ce cas, plus on a de voisins proches de z_j, plus la valeur de $\sum w_{ij} z_i$ est grande : on ajoute dans le modèle une somme pondérée des valeurs des voisins et non une moyenne pondérée.

- Les régressions avec poids spatial (GWR) utilisent également une matrice de poids spatiaux. On permet ici aux coefficients β du modèle de varier en fonction de la localisation, afin d'adapter localement le modèle aux variations spatiales locales : ces modèles visent à estimer localement les paramètres de la régression.

La standardisation sur les lignes de la matrice de distances W (chaque poids étant divisé par la somme des poids de sa ligne) peut également être utilisé dans le calcul des indices de Moran ou de Geary, ce qui revient à prendre comme indice global la moyenne arithmétique des indices locaux.

2. La standardisation spatiale

Dans un ensemble de points quelconques, les couples de points ne sont pas répartis de façon constante en fonction de leur distance : en général le nombre de couples de points augmente avec la distance, jusqu'à atteindre un maximum puis diminuer. Cette distribution statistique des distances entre les points (appelées dans la suite inter-distances) dépend de la distribution spatiale des points.

A titre d'exemple, lorsque les points sont distribués de façon indépendante et uniforme dans un disque de rayon R, le nombre de points de trouvant dans une couronne de rayon $[R, R+d]$ augmente de façon linéaire avec ce rayon, contrairement à la dimension 1 où il reste constant (Figure 1) :

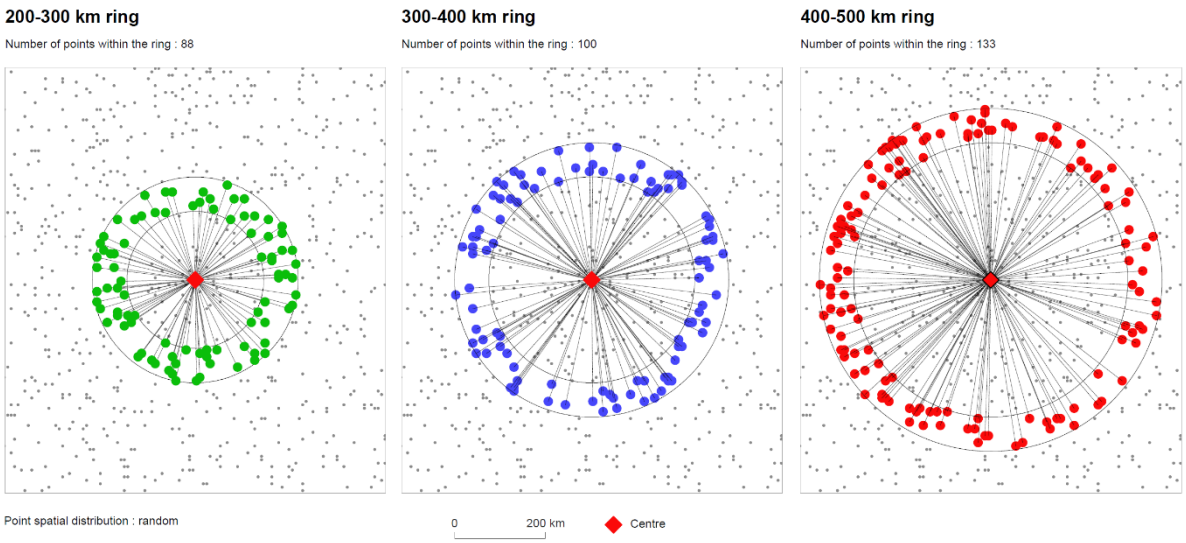


Figure 1. Le nombre de points dans une couronne de même largeur augmente proportionnellement au rayon, pour des points indépendamment et uniformément distribués dans l'espace de dimension 2.

La distribution des distances des segments formés par tous les couples de points se trouvant dans D s'exprime par la fonction de répartition suivante [18, 19] (Figure 2) :

$$\forall d \in]0;2R], \quad f(d) = \frac{4d}{\pi R^2}\left(\arccos\left(\frac{d}{2R}\right) - \frac{d}{2R}\sqrt{1-\left(\frac{d}{2R}\right)^2}\right) \tag{6}$$

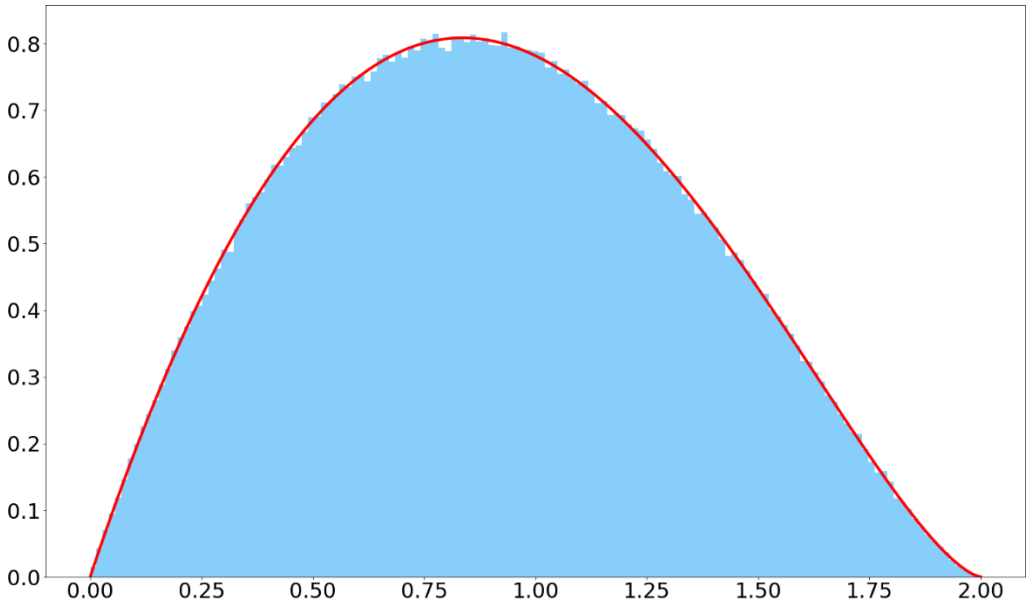


Figure 2. Distribution des inter-distances dans le cercle unité (R=1) pour un ensemble de points indépendamment et uniformément répartis dans D. En rouge, la courbe de la fonction de densité ; en bleu, la distribution des inter-distances dans un nuage de points simulé à partir d'un modèle de Poisson homogène (densité $\rho = 1500$ dans le cercle unité).

Ce constat très simple n'est pas pris en compte dans les méthodes courantes d'analyse spatiale qui utilisent la somme ou la moyenne des valeurs pondérées et des distances pour calculer les poids spatiaux pour caractériser et analyser la dépendance spatiale. La distribution des inter-distances n'étant pas constante, les calculs basés sur des sommes ou des moyennes sur tous les couples de points favorisent les valeurs des couples de points dont les distances sont les plus fréquentes, alors que la dépendance spatiale ne devrait être caractérisée que par une fonction de la distance. Lorsqu'un calcul implique une somme ou une moyenne sur les couples de poids, il faut éliminer l'influence de la distance sur le nombre d'inter-distances dans le calcul pour ne pas sur-représenter ou sous-représenter les inter-distances de certains couples de points, alors que la distance est justement la variable explicative principale. Le poids spatial W utilisé dans le calcul ne traite pas ce problème, car il est construit pour modéliser la dépendance spatiale et non le fait que certaines inter-distances sont systématiquement plus représentées que d'autres dans le calcul de l'indice ou de l'estimation.

La plupart du temps, la distance d'influence $dmax$ utilisée dans le calcul du poids spatial est inférieure à la distance pour laquelle le nombre d'inter-distances atteint un maximum. Dans ce cas, le nombre d'inter-distances passe de 0 à $dmax$, et il est très probable que l'influence du poids spatial dans le calcul (qui favorise en général les courtes distances) soit annulée par la non prise en compte la distribution des inter-distances.

Dans cet article, nous proposons donc une amélioration des méthodes qui utilisent des poids spatiaux faisant appel à la distance et à une somme ou une moyenne dans le calcul. Cette amélioration s'apparente à une « standardisation spatiale », à l'image de la standardisation classique en une dimension (comme par exemple la standardisation sur l'âge). Elle est différente de la standardisation sur les lignes de la matrice des poids spatiaux qui ne résout pas le problème de la distribution des

inter-distances, puisque chaque ligne correspond à un indice local et fait face au même problème (chaque calcul d'indice local favorise les valeurs des points distants).

3. Méthodes

Nous proposons d'ajuster le calcul des indices ou des estimations en apportant une correction (appelée dans la suite SD-correction) afin de supprimer l'influence de la distribution statistique des inter-distances sur les calculs. Pour cela, nous proposons d'ajouter un second poids pour chaque couple de points (P_i, P_j) dans le calcul de la somme. Ce second poids w'_{ij} correspond pour chaque inter-distance $d(P_i, P_j)$ à l'inverse de l'influence relative de l'inter-distance dans l'ensemble des toutes les inter-distances prises en compte dans le calcul. Il est donné par l'inverse $1/p_{ij}$ de la probabilité p_{ij} de l'inter-distance dans l'ensemble de toutes les inter-distances prises en compte dans le calcul. Pour chaque inter-distance, il est calculé à partir de la fonction f de distribution des inter-distances.

Cette distribution peut être donnée par la fonction de répartition des interdistances lorsque cette fonction est connue (comme par exemple le cas cité plus haut, où la distribution spatiale du nuage de point est définie par une distribution spatiale connue).

Lorsque la fonction de répartition des inter-distances est inconnue, pour chaque situation donnée nous proposons d'approcher cette distribution soit par une fonction constante par morceau, en calculant le nombre relatif d'inter-distances $\frac{N(k)}{N}$ dans chaque intervalle $[d, d+h[$ avec un pas h fixé et $d = kh$ $(k \in \mathbb{N})$, d variant entre 0 et $dmax$ (borne maximale des inter-distances à considérer), où N est le nombre total d'inter-distances entre 0 et $dmax$. On effectue le calcul soit par interpolation entre les points $(kh + h/2, N(k))$ avec une fonction affine par morceaux, soit par interpolation par noyau, avec comme noyau une fonction gaussienne $\frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{d}{h})^2}$ d'écart-type h. Le pas h permets de paramétrer l'influence de la standardisation sur la distance sur l'ensemble du calcul. Le poids w_{ij} d'un couple de points (P_i, P_j) est donc modifié en le divisant par une approximation de la fonction de densité $f(d(P_i, P_j))$.

Par exemple, dans le premier cas l'indice de Moran corrigé sera :

$$\hat{I}_{Moran} = \frac{1}{S'} \sum_{i,j} w_{ij} w'_{ij} \left(\frac{X_i - m}{\sigma}\right)\left(\frac{X_j - m}{\sigma}\right) \quad (7)$$

où w_{ij} est le poids spatial pour la dépendance spatiale, $w'_{ij} = 1/p_{ij}$ et p_{ij} la probabilité de l'inter-distance $d(P_i, P_j)$, donnée par l'approximation de la fonction de densité de probabilité $f(d(P_i, P_j))$ calculée comme indiqué plus haut. S' est la somme des poids $w_{ij} w'_{ij}$ $(S' = \sum_{i,j} \frac{w_{ij}}{p_{ij}})$.

L'espérance de l'indice de Moran SD-corrigé $\hat{I}$ est égale à l'espérance de l'indice de Moran original I, puisque cette espérance ne dépend pas des poids.

Avec les même notations, l'indice SD-corrigé de Geary a pour expression :

$$\hat{I}_{Geary} = \frac{1}{S'} \sum_{i,j} w_{ij} w'_{ij} \left(\frac{X_i - m}{\sigma} - \frac{X_j - m}{\sigma}\right)^2 \quad (8)$$

La SD-correction appliquée au calcul des indices d'autocorrélations et aux estimation spatiales par noyau a été implémentée dans SavGIS, un logiciel SIG en téléchargement gratuit (www.savgis.org). Ce logiciel a été utilisé pour développer l'exemple qui suit.

4. Exemple

A titre d'exemple, nous avons appliqué la SD-correction à l'analyse spatiale des résultats de l'élection présidentielle qui a eu lieu en France en avril 2017. La variable analysée est le pourcentage de votes remportés par le candidat Emmanuel Macron au second tour en France métropolitaine, agrégés par canton électoral (la France métropolitaine est divisée en 1971 cantons électoraux). Les données utilisées sont disponibles en libre accès sur le site Internet du gouvernement français (https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-23-avril-et-7-mai-2017-resultats-du-2eme-tour-1/).

Comme cela a été constaté dans de nombreux pays, le comportement électoral montre généralement une certaine continuité dans l'espace, bien que des différences notables puissent exister entre des unités spatiales voisines. Cette variable est donc bien adaptée à l'analyse et au calcul de l'autocorrélation spatiale. La figure 3 met en évidence un contraste entre les villes, largement en faveur d'Emmanuel Macron, et les zones rurales, en particulier le nord-est de la France et la côte méditerranéenne où l'extrême droite a obtenu le plus grand nombre de voix. Sur le côté droit de la figure 3, les graphiques fournissent respectivement la distribution des distances entre les plus proches voisins (la distance moyenne entre cantons adjacents est de 18 km, centroïde à centroïde), et la distribution des inter-distances (distances entre chaque couple de points).

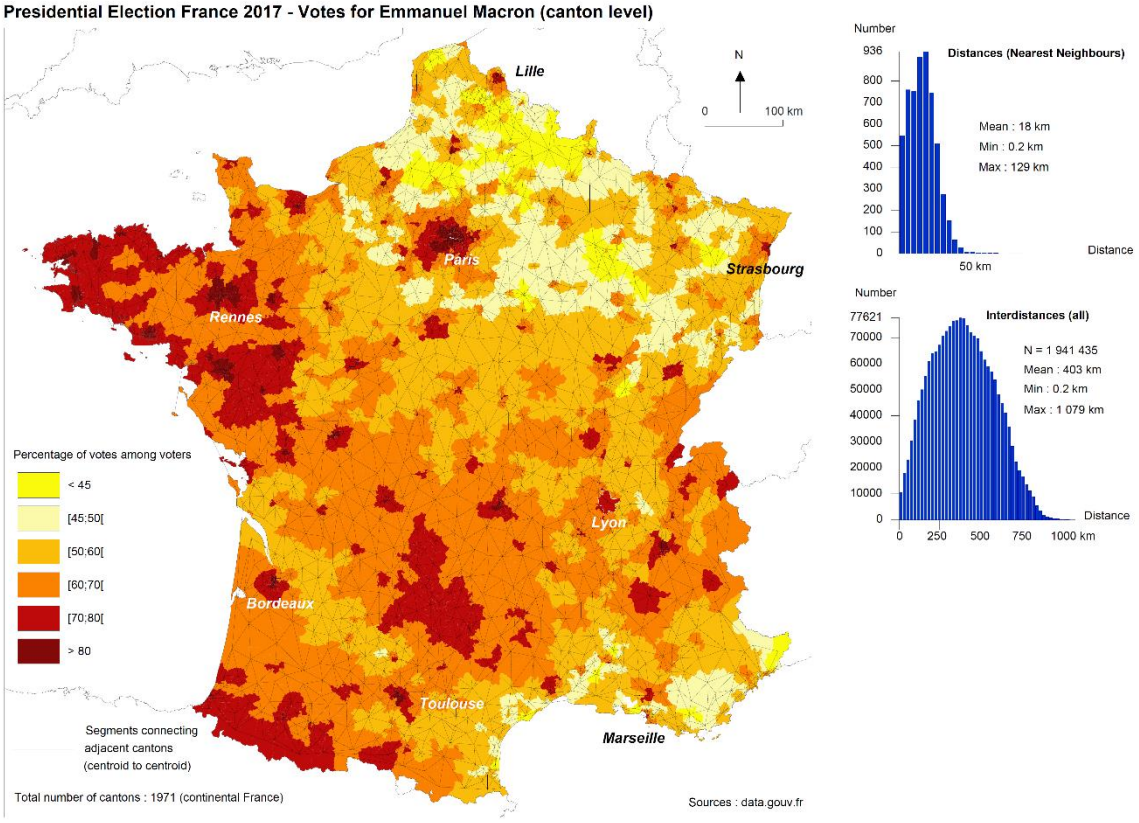


Figure 3. Votes pour Emmanuel Macron (%) au second tour de l'élection présidentielle en France, au niveau des cantons (2017) (source : data.gouv.fr and Institut Géographique National-IGN).

4.1. Indices autocorrélation spatiale

A partir de ces pourcentages, nous avons calculé les indices d'autocorrélation spatiale de Moran et de Geary, avec et sans SD-correction. Le semi-variogramme du pourcentage de votes en faveur d'Emmanuel Macron montre une influence spatiale inférieure à 250 km (Figure 4). Nous avons calculé les indices de Moran et de Geary avec $dmax$ variant de 25 à 250 km (figure 5, tableau 1).

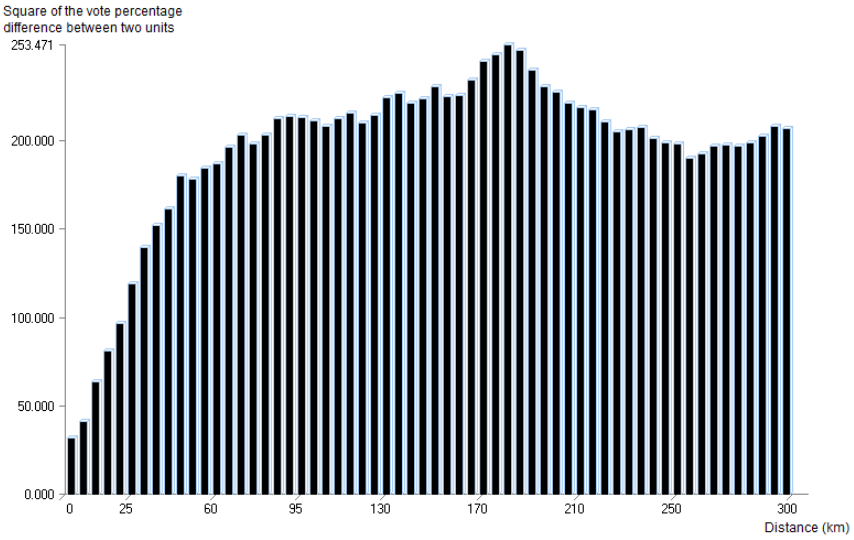


Figure 4. Semi-variogramme du pourcentage de votes en faveur d'Emmanuel Macron.

La figure 5 et le tableau 1 montrent que la valeur de l'indice de Moran ou de Geary avec SD-correction (en vert) montre une autocorrélation spatiale plus forte que celle de l'indice non corrigé (en jaune). La figure 5 indique également que les valeurs des deux indices (Moran ou Geary), corrigées et non corrigées, montrent une autocorrélation décroissante constante lorsque la largeur de l'influence maximale $dmax$ augmente. C'est logique, car la dépendance spatiale des valeurs entre les unités spatiales diminue lorsque la distance augmente, et le fait de considérer des couples plus éloignés a tendance à réduire la moyenne globale pondérée.

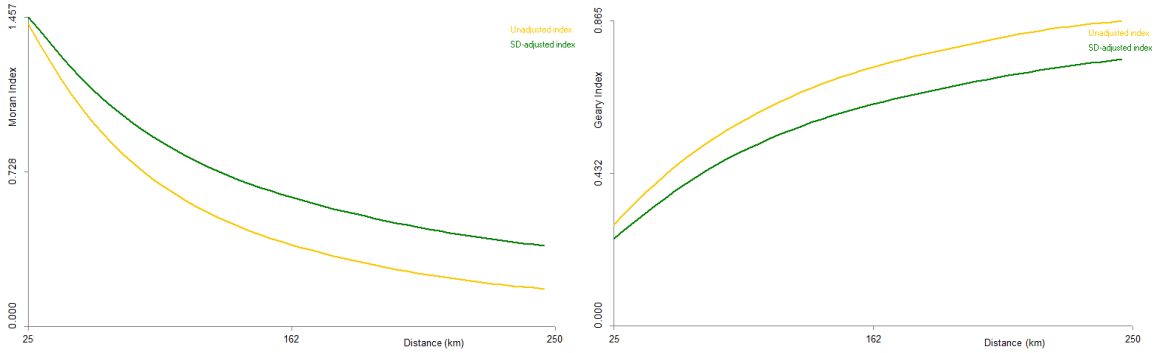


Figure 5. Indices de Moran (à gauche) et de Geary (à droite) avec une influence maximale $dmax$ variant de 25 à 250 km. En jaune sans SD-correction, en vert avec SD-correction.

Table 1. Valeurs des indices de Moran corrigés et non-corrigés pour des valeurs $dmax$ croissantes.

Influence max. (km)	Nombre de couples	Indice de Moran non corrigé	Z-Score	Ecart-type	Indice de Moran corrigé	Z-Score	Ecart-type
25	12,804	1.43	138.78	0.0104	1.46	33.46	0.0433
50	34,012	1.01	165.45	0.0062	1.14	45.47	0.0251
75	62,930	0.73	156.81	0.0047	0.92	89.67	0.0101
100	101,544	0.55	145.10	0.0036	0.76	128.16	0.0059
150	206,304	0.35	134.88	0.0025	0.57	149.73	0.0038
200	337,288	0.24	124.62	0.0019	0.45	150.47	0.0029

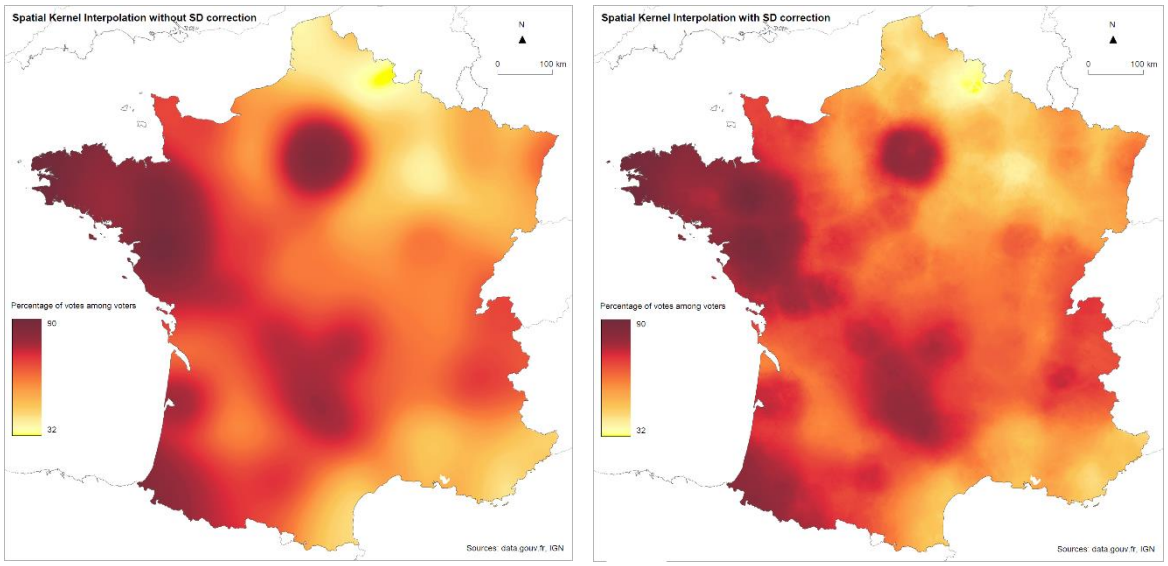
250	484,394	0.18	113.65	0.0016	0.37	152.98	0.0025
300	644,016	0.14	110.83	0.0013	0.32	153.41	0.0020

La signification statistique du rejet de l'hypothèse nulle (H0) d'absence d'autocorrélation spatiale (fournie ici par le Z-Score correspondant à la valeur de l'indice observé) est essentielle pour conclure à la présence d'autocorrélation spatiale. Dans cet exemple, tous les indices, corrigés et non corrigés, montrent une très forte probabilité pour la présence d'autocorrélation spatiale, comme on peut le voir dans le tableau 1, avec pour toutes les valeurs $dmax$ testées, des valeurs très élevées du Z-Score, ce qui correspond à des p-value très faibles. Nous notons également dans cet exemple que la significativité de l'indice SD-corrigé augmente lorsque la distance d'influence $dmax$ augmente, alors que la significativité de l'indice non corrigé diminue (à partir d'une largeur de 75 km). Nous pouvons également noter que la variance est plus élevée pour l'indice de Moran corrigé que pour l'indice original.

Si les valeurs sont assignées au hasard aux unités géographiques pour détruire l'autocorrélation spatiale, les indices non corrigés et SD-corrigés présentent des valeurs similaires : la correction n'agit qu'en présence d'une autocorrélation spatiale.

4.2. Interpolation spatiale par noyau

La figure 6 montre que pour le même noyau (dans ce cas, une fonction gaussienne avec h = 200 km), la SD-correction augmente la précision de l'interpolation. La SD-correction donne beaucoup plus de détails et le résultat est moins lissé. Sans correction, il y a plus de couples de points associés à de grandes distances dans la limite de h, et ces couples de points ont une plus grande influence dans le calcul. Ceci réduit l'influence des couples de points moins fréquents et conduit à des résultats beaucoup plus moyennés, basés sur des couples de points de plus grande distance. Il produit une surface de tendance beaucoup plus lisse mais moins précise.



(**a**) (**b**)

Figure 6. Interpolation spatiale par Noyau (fonction gaussienne, h = 200 km) appliquée au pourcentage de vote en faveur d'Emmanuel Macron au second tour de l'élection présidentielle en France (2017) : (**a**) carte de gauche sans SD-correction ; (**b**) carte de droite avec SD-correction.

5. Discussion et conclusion

Cet article passe en revue l'un des fondements de l'analyse spatiale : l'évaluation de la dépendance spatiale. Bien que la diminution des poids spatiaux avec l'augmentation de la distance ait déjà été largement analysée et discutée dans la littérature, la distribution inégale des couples de points en

fonction de la distance a jusqu'ici été négligée. Pourtant, cette distribution inégale des couples de points a un impact direct sur les calculs utilisés dans un grand nombre de méthodes d'analyse spatiale, telles que les indices d'autocorrélation spatiale, les méthodes d'interpolation par noyau ou les méthodes de modélisation spatiale. La distribution statistique des inter-distances n'étant pas constante, toutes les méthodes qui reposent sur le calcul d'une somme ou d'une moyenne de valeurs de couples de points favorisent les valeurs des couples de points dont les distances sont les plus représentées, alors que la dépendance spatiale ne devrait être évaluée qu'en fonction de la distance. Lorsqu'un calcul implique une somme ou une moyenne sur les couples pondérés, l'influence de la distance qui découle de la distribution inégale des interdistances doit être corrigée pour ne pas sur- ou sous-représenter certaines inter-distances, car la distance est précisément la principale variable explicative pour la dépendance spatiale. Pour résoudre ce problème, nous avons introduit le concept de « standardisation spatiale » et un nouveau poids $w'_{ij} = 1/p_{ij}$, où p_{ij} est la probabilité de l'inter-distance $d(P_i, P_j)$, donnée par la fonction de densité des inter-distances.

Logiquement, et comme le montre bien l'exemple détaillé ci-dessus, l'effet de la correction devient de plus en plus important lorsque la distance d'influence augmente et que le nombre d'inter-distances impliquées dans le calcul augmente. En effet, dans le calcul des indices non corrigés, l'augmentation relative du nombre d'inter-distances longues par rapport aux inter-distances courtes réduit l'influence de la dépendance spatiale, puisque le poids spatial (qui modélise la dépendance spatiale entre deux objets) diminue avec la distance. Lorsque la distance augmente, le nombre de couples de points de plus grande distance augmente, ainsi que leur influence dans le calcul. L'effet de la dépendance spatiale dans le résultat du calcul est donc réduit. La SD-correction vise à équilibrer cette influence. L'indice ou l'estimation corrigé montre des valeurs d'autocorrélation plus fortes que l'indice ou l'estimation non corrigé, en donnant plus de poids aux inter-distances moins fréquentes dans le calcul, et donc, en général, aux couples de courtes distances - précisément celles qui montrent, en présence de dépendance spatiale, la plus forte corrélation entre leurs valeurs. La SD-correction renforce ainsi l'objectif des indices d'autocorrélation (capturer et mesurer l'autocorrélation spatiale) lorsque le calcul implique une somme ou une moyenne pondérée de valeurs de couples de points.

Dans le cas de l'indice de Moran, nous avons également vu dans notre exemple que la SD-correction augmentait la variance de l'indice. La variance de l'indice de Moran dépend des poids spatiaux [10]. La SD-correction ajuste les poids en rééquilibrant la valeur relative des poids en fonction de la distribution des inter-distances. Elle augmente donc la variance des poids spatiaux, ce qui se reflète dans la variance de l'indice lui-même.

Les estimations ou indices corrigés peuvent être plus sensibles que les estimations non corrigées aux valeurs des inter-distances les plus courtes, car le poids corrigé de ces inter-distances est le produit du poids spatial (en général plus élevé pour les inter-distances courtes afin de saisir l'autocorrélation spatiale) et du poids de la correction (qui dépend de la distribution spatiale des points, mais qui est presque toujours supérieur pour les inter-distances courtes et les inter-distances longues). Cette remarque sur la variance montre également que la correction proposée renforce la capacité des indices corrigés à mesurer l'autocorrélation spatiale, en donnant plus de poids aux inter-distances les plus courtes dans le résultat, mais entraînant une augmentation de la variance.

En conclusion, cet article montre qu'il est important de mettre en œuvre la SD-correction pour toutes les méthodes, modèles et estimations qui impliquent des calculs d'autocorrélation spatiale basés sur des sommes ou des moyennes de valeurs pondérées en fonction de la distance.

Contributions des auteurs : Conception, analyse, méthode et logiciels : Marc Souris ; validation, exemple : Florent Demoraes ; rédaction et préparation du manuscrit original, correction et mise en forme : Marc Souris et Florent Demoraes.

Financement : Ces travaux n'ont pas reçu de financements extérieurs.

Conflits d'Intérêt : pas de conflits d'intérêt déclarés.

Références

1. Tobler W. A computer movie simulating urban growth in the Detroit region. *Econ. Geography Suppl.* **1970**, *46*, 234–240. [CrossRef]
2. Shabenberger, O.; Gotway, C. *Statistical Methods for Spatial Data Analysis*; Chapman & Hall: London, UK, 2005.
3. Souris, M. *Epidemiology and Geography. Principles, Methods and Tools of Spatial Analysis*; Wiley-ISTE: London, UK, 2019. *Epidemiologie et Géographie, Principes, Méthodes et Outils de L'analyse Spatiale*; ISTE: London, UK, 2019, pour la version française.
4. Moran, P. The interpretation of statistical maps. *J. R. Stat. Soc. Ser. B* **1948**, *10*, 243–251. [CrossRef]
5. Geary, R.C. The contiguity ratio and statistical mapping. *Inc. Stat.* **1954**, *5*, 115–145. [CrossRef]
6. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
7. Getis, A.; Ord, J.K. The analysis of spatial association by use of distance statistic. *Geogr. Anal.* **1992**, *24*, 189–206. [CrossRef]
8. Fotheringham, S.; Rogerson, P.A. *The Sage Handbook of Spatial Analysis*; Sage: London, UK; Los Angeles, CA, USA, 2009.
9. Cliff, A.D.; Ord, J.K. *The Problem of Spatial Autocorrelation*; Scott, A.J., Ed.; Studies in Regional Science; Pion: London, UK, 1969; pp. 25–55.
10. Cliff, A.D.; Ord, J.K. *Spatial Processes: Models and Applications*; Pion Limited: London, UK, 1981.
11. Upton, G.J.G.; Fingleton, B. *Spatial Data Analysis by Example*; Wiley: New York, NY, USA, 1985.
12. Anselin, L.; Bera, A.K. Spatial dependence in spatial regression model, with an introduction to spatial econometrics. In *Handbook of applied Economic Statistics*; Ullah, A., Giles, D.E., Eds.; Marcel Decker: New York, NY, USA, 1988; pp. 237–289.
13. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **1967**, *27*, 209–220. [PubMed]
14. Getis, A.; Ord, J.K. Local spatial statistics: An overview.: In *Spatial Analysis: Modeling in A GIS Environment*; Longley, P., Batty, M., Eds.; John Wiley & Sons: New York, NY, USA, 1996; pp. 261–277.
15. Droesbeke, J.J.; Lejeune, M.; Saporta, M. *Analyse Statistique des Données Spatiales*; Technip: Paris, France, 2006.
16. Bowman, A.W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis*; Oxford University Press: London, UK, 1997.
17. Dormann, C.; McPherson, J.; Araújo, M.; Bivand, R.; Bolliger, J.; Carl, G.; Davies, R.G.; Hirzel, A.; Jetz, W.; Kissling, W.D.; et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* **2007**, *30*, 609–628. [CrossRef]
18. Alagar, V.S. The distribution of the distance between random points. *J. Appl. Probab.* **1976**, *13*, 558–566. [CrossRef]
19. Lellouche, S.; Souris, M. Distribution of distances between elements in a compact set. Unpublished, manuscript in preparation.