Chapitre 11
Statistiques spatiales
Représenter et analyser les organisations dans l'espace géographique

Hélène MATHIAN
Environnement Ville et Société - EVS, UMR 5600, CNRS, Univ. de Lyon, Lyon

1 Introduction : méthodes statistiques et données géographiques

L'objet de ce chapitre est de présenter des méthodes statistiques appliquées à des données géographiques, c'est-à-dire à références spatiales. La terminologie associée à cet ensemble de méthodes n'est pas forcément consensuelle. Elles réfèrent toutes à un traitement, généralement statistique, de données numériques ayant une composante géographique et, résultent le plus souvent d'un dialogue méthodologique entre géographie et statistiques. Sous l'angle de la géographie, elles se situent à l'intérieur de catégories assez larges telles que les méthodes de *géographie quantitative* ou encore d'*analyse spatiale** qui désigne autant les méthodes que des modèles théoriques sous-jacents. Si l'on entre par les statistiques, on trouve différentes désignations, allant de l'*analyse de données géographiques*, les *statistiques spatiales*, les méthodes *géostatistiques*… faisant référence à des d'approches méthodologiques différentes : descriptives et exploratoires pour les méthodes d'analyse de données géographiques, confirmatoires et inférentielles pour les méthodes de *statistiques spatiales* et les *géostatistiques* (Fotheringham et Rogerson 2008). La diversité des usages de ces méthodes relève souvent d'habitudes disciplinaires différentes, comme par exemple, dans le domaine de l'écologie (Wiegand et Moloney 2013, Dale et Fortin 2014), de l'archéologie (Hodder et Orton 1976, Conolly et Lake 2006, Nakoinz et Knitter 2016), voire de l'épidémiologie (Pfeiffer, Dirk *dir.* 2009, Lawson *et al.* 2016). Quelles que soient leurs désignations, l'ensemble de ces méthodes constituent une boîte à outils méthodologique riche dans laquelle on peut aller puiser selon un type d'approche ou un autre, et dont les fonde-

ments conceptuels permettent d'interroger les structures spatiales et les processus qui conduisent à ces structures, autant méthodologiquement que thématiquement. Ainsi ces méthodes peuvent être utilisées pour « confirmer » un modèle ou pour « explorer » des données, en portant autant attention à la signification thématique qu'à la significativité statistique (Gould 1970).

Avant d'aborder ces méthodes, il nous parait important d'introduire certaines spécificités des données dites géographiques, c'est-à-dire des données qui font référence à l'espace, ou encore dîtes géoréférencées. Une première spécificité de ce type de données concerne la manière dont on conçoit la spatialité du phénomène qu'elles décrivent. Ainsi le phénomène spatial pourra être conçu comme :

- Discret : c'est-à-dire ayant une empreinte spatiale constituée de régions ou zones (aire culturale, parcelles…) ou caractérisée par un ensemble d'objets localisés (des haches (*cf.* 2.2.7) ou des établissements* (*cf.* 2.2.4))
- Continu : c'est-à-dire défini en tout point de l'espace (température, pluviométrie, teneur en minéraux du sol… *cf.* 2.2.6).

Cette première différenciation met en perspective au niveau conceptuel, une vue « objet » (phénomène discret) et une vue « champs » (phénomène continu) des phénomènes observés, qui correspondent aussi à deux modèles informatiques différents, à savoir respectivement **vecteur** et **raster**[1].

Nous nous situons ici dans une perspective méthodologique essentiellement « objet », même si certaines de ces méthodes, nous le verrons, pourront être associées à des phénomènes continus. En statistique, on appelle *données* un ensemble de mesures* décrivant des individus*. Lorsque l'on travaille sur des phénomènes spatiaux, les individus que l'on décrit désignent généralement des *entités spatiales.* Elles sont décrites par des mesures relevant de deux logiques. On dit qu'elles comportent deux composantes. L'une, classique, décrit la sémantique (type d'habitat, occupation du sol des parcelles…) et est notée généralement [Z] (elle peut être une valeur ou un vecteur de données). On l'appelle aussi composante attributaire. L'autre composante est géométrique et décrit l'empreinte spatiale de l'objet (elle est notée [XY]). [XY] pourra désigner un ou plusieurs points, une ou plusieurs lignes, ou encore un ou plusieurs polygones (Barge, Saligny 2011, Wheatley, Gillings 2002).

Les traitements de données géographiques vont donc avoir cette spécificité de combiner les deux composantes. Nous discutons dans un premier temps de comment la dimension spatiale à la fois, contrevient aux conditions ou aux hypothèses imposées par les méthodes de statistiques, et contient intrinsèquement des potentialités pour décrire et analyser les organisations (2). Nous illustrerons ensuite deux approches pour analyser les organisations, fondées sur des méthodes de statistiques spatiales : l'analyse de l'autocorrélation spatiale (3) et les méthodes associées à l'analyse des semis de points (4). Enfin nous aborderons dans une dernière partie (5) les méthodes

[1] Le mode vecteur ou vectoriel défini un objet géométrique composé de nœuds (des points dans l'espace) et des formules mathématiques pour calculer les arcs (des lignes) qui relient ces nœuds entre eux. Le mode raster est une image numérique composée d'un ensemble de cellules (pixels).

qui permettent de passer d'observations ponctuelles à des surfaces continues que ce soit par interpolation, lissage ou calcul.

2 Les spécificités des données spatiales

Lorsque l'on introduit les méthodes statistiques appliquées aux données géographiques, on utilise souvent l'adage « le spatial est spécial ». Il est effectivement important en préambule de revenir sur les spécificités de ces données car cette « spécialité » vient enfreindre des propriétés nécessaires à l'utilisation des méthodes statistiques dans un contexte inférentiel (*cf.* 8). Par ailleurs, la composante spatiale peut être mobilisée de multiples formes que nous présentons ici et qui seront utiles pour la compréhension des méthodes.

2.1 Les différentes représentations géométriques d'un objet

D'un point de vue formel, la composante géométrique désigne les géométries permettant de représenter les objets géographiques : elles peuvent être des points, des lignes, des polygones (Longley *et al.* 2015). Toutefois il s'agit bien de *représentation*, et, selon le point de vue, cette représentation peut être généralisée ou simplifiée. Par exemple le calcul de la superficie d'un habitat* se fera avec le polygone associé à son dessin sur le sol, le plus détaillé et proche de la « réalité » ; alors que pour calculer la distance de cet habitat à la route, on représentera l'habitat par un point (par exemple l'entrée, ou le centre de gravité du polygone) et on calculera la distance entre ce point (représentation de l'habitat) et le point de la route le plus proche. La composante géométrique peut ainsi prendre plusieurs formes.

En outre, l'espace peut être pris en compte selon les niveaux ou échelons géographiques qu'il porte de manière intrinsèque : cela définit la granularité* à laquelle on travaille (*cf.* 2.2.2). On peut ainsi définir tout un système d'entités spatiales de différents niveaux géographiques ou de différentes portées, qui illustrent l'organisation hiérarchiques soit des systèmes d'appropriation de l'espace soit des pratiques. On peut par exemple partir de données décrivant un ensemble d'habitats, puis agréger les habitats et leurs descriptions au sein d'agglomérations par exemple. On obtient ainsi des objets ayant une nouvelle empreinte spatiale, et dont certains attributs vont être propres à ce niveau (surface de l'agglomération par exemple) et d'autres vont provenir de la description des habitats qui les contiennent (nombre total, distribution des habitats par type de matériaux de construction, superficie occupée, …)

Les réponses à ces questions de représentation géométrique d'une part et de définition du niveau d'analyse d'autre part, doivent être guidées par une réflexion conjointe :

- Méthodologique - quelle type de géométrie est nécessaire pour quelle méthode ?
- Et ontologique - quel sens donner aux résultats de l'analyse de cette donnée - (*cf.* chapitre introduction)

2.2 Hypothèses statistiques et données géographiques

Les propriétés liées à la composantes géométriques des **données géographiques** contreviennent à deux règles fondamentales des conditions d'applications des méthodes statistiques.

- Les observations* ne peuvent être considérées comme *identiquement distribuées*. L'espace géographique n'est pas neutre en soit : il existe des pleins et des vides de ressources susceptibles d'influencer des valeurs et ces différenciations impactent les valeurs observées. On dit qu'il existe des « effets de site ».
- L'hypothèse d'*indépendance* des observations* n'est pas respectée : les données géographiques sont liées par le phénomène d'autocorrélation spatiale : ce qu'il se passe en un lieu* a une influence sur ce qu'il se passe « autour ». Les valeurs proches ont plus de chance d'être proches que des valeurs éloignées. Ce sont les « effets de situation ».

Au-delà de ces « effets spatiaux » susceptibles de venir biaiser les résultats, le contexte même d'analyse est à réinterpréter en fonction des spécificités des données spatiales, qui ne sont pas des échantillons* et quasiment jamais aléatoires (P. Gould). On se trouve ainsi devant une forme de paradoxe dans le cadre d'analyse géographique de données archéologiques : les géographes étudient l'ensemble d'un territoire, et compte-tenu justement des relations entre les entités, ils envisageraient difficilement de tirer un échantillon aléatoire de ces données qui donneraient une vision partielle du phénomène. En archéologie le plus souvent tout l'espace n'est pas renseigné, et les données sont un échantillon d'une population* que l'on ne connaît pas. Mais cet échantillon ne peut généralement être considéré comme aléatoire[2]. Dans ces conditions, que ce soit en géographie ou en archéologie, on ne peut faire confiance à une application classique des méthodes statistiques et aux résultats pris d'un point de vue inférentiel.

Concernant les effets liés à la spécificité des données spatiales, on ajoutera d'autres points de vigilance à'intégrer dans la chaine de traitement et d'interprétation de données géographiques :

- La première relève de ce que l'on appelle les *effets de bord** : la première étape d'une analyse consiste à délimiter spatialement la zone d'étude, déterminant par là même la population statistique considérée. Or, les systèmes spatiaux ne sont pas des systèmes fermés. Quel que soit le phénomène analysé, il comporte des éléments mobiles (pollen, habitants…), et leurs mobilités ne s'arrête généralement pas à la limite de la zone d'étude. Par exemple, le vent disperse le pollen bien au-delà des limites de la zone occupée par un type de végétal. Les frontières

[2]Les contextes archéologiques résultent de processus taphonomique, c'est-à-dire d'un ensemble de phénomènes naturels et anthropiques qui affectent au cours du temps (ou de l'enfouissement) l'intégrité des objets : leur nature et leur relations spatiales. Cet impact peut même se traduire par une disparition totale des traces de l'objet (dégradation organo-chimique ou subtilisation humaine par exemple). Les analyses taphonomiques peuvent mettre en évidence la logique d'altération des objets archéologiques et la distribution des artefacts produits (par exemple celle des objets lithiques issus des travaux de taille de silex, tri, balayage…ou encore celle liée à l'enfouissement et au déplacement de vestiges dans une zone de colluvionnement).

les plus strictes n'empêchent pas la mobilité. Ainsi, plus on se rapproche des « bords » de la zone d'étude, plus le voisinage des entités qui y sont localisées sera tronqué. Ceci aura un impact sur les résultats des méthodes s'appuyant sur ces voisinages.

- Les deux autres sont généralement identifiés par le MAUP (Modifiable Areal Unit Problem) (Openshaw 1984) qui définit deux effets liés à des caractéristiques déjà abordée ci-dessus (2.1) à savoir : *l'effet d'échelle* lié au fait qu'une même information peut être analysée à différents niveaux géographiques ; et *l'effet de maillage* qui désigne le fait qu'un même niveau géographique peut être placé de multiples manières (selon par exemple le point d'origine d'un maillage régulier). Si ces effets définissent un « problème » du point de vue statistique c'est qu'ils sont susceptibles de contrevenir à la notion de robustesse* des résultats qui est associé à la validation des résultats en statistiques. En effet, une relation calculée sur des données agrégées (superficie des habitats et niveau hiérarchique par exemple) peut être significative à un échelon et ne plus l'être à une autre, ou même être simultanément positive et négative selon le maillage considéré. Le point de vue géographique défend que les échelons et maillages ne sont que des révélateurs d'agencements et de niveaux d'organisations différents. Cette variabilité des résultats en fonction du maillage considéré n'apparaît pas comme un problème puisqu'une variable* calculée à différents échelons constitue sémantiquement des observables* différents. À même échelon géographique, deux maillages différents vont révéler des agencements spatiaux différents. La richesse de l'analyse spatiale est de pouvoir combiner ces différents points de vue et d'enrichir la compréhension du phénomène. Le « P » de MAUP peut être celui de *potentialités* et intégrer la richesse qu'apportent ces complémentarités de point de vue (Grasland et Madelin 2006).

Les méthodes de statistiques spatiales qui sont présentées ci-dessous s'affranchissent des questions soulevées par ces différentes caractéristiques en raisonnant en termes de **processus spatial** : une organisation spatiale (agencement des valeurs dans l'espace) est une réalisation d'un processus spatial. En posant des hypothèses sur ce processus il est alors possible d'interroger l'agencement observé, au regard de l'ensemble des configurations possibles produites par un processus. Généralement on fait l'hypothèse que le processus est indépendant de l'espace c'est-à-dire qu'il vérifie les propriétés (1) tout événement/valeur a la même probabilité* d'apparaître en tous points de l'espace et (2) l'apparition d'un événement/valeur en un lieu n'a pas d'incidence sur les autres valeurs de l'espace. L'agencement spatial résultant d'un processus aléatoire est dit aléatoire. On peut construire des statistiques permettant de tester la probabilité d'apparition de l'agencement observé sous cette hypothèse de processus aléatoire indépendant de l'espace. Ce que l'on teste alors n'est rien d'autre que l'influence des deux facteurs décrits ci-dessus, à savoir *l'effet de site* et *l'effet de situation* qui conditionnent des processus non aléatoires tels que la concentration, la compétition, la diffusion, la contagion…

2.3 Formalisation des relations dans l'espace : distance, voisinage et interaction spatiale

Parmi les méthodes que nous présentons on peut distinguer deux niveaux de spécificités : ce sont soit des méthodes dédiées à des données spatiales, soit des méthodes plus génériques mais qui peuvent être spécifiées pour des données spatiales. Leur point commun est qu'elles intègrent dans leur formalisme une expression de ce qui lie ces données entre elles, à savoir leur proximité spatiale. Elles intègrent le fait, ainsi que l'a énoncé W. Tobler[3] : "*everything is related to everything else, but near things are more related than distant things*" (1970).

Nous proposons ici de décliner les différents aspects de ces relations spatiales d'un point de vue formel :

- La première formalisation est celle de la notion de distance. Dans les développements statistiques la distance est entendue au sens mathématique du terme et la distance la plus utilisée est la distance euclidienne*. Lorsqu'on analyse l'espace, la notion de distance fait référence le plus souvent à des distances empiriques, pratiquées. Ce peut être des distances calculées sur un réseau, mais aussi des temps ou des coûts de parcours (distance temps, distance coût…). Ce que l'on appelle communément distance en géographie ne vérifient pas nécessairement les propriétés d'une distance mathématique : l'espace géographique n'est pas isotrope et les formes et niveaux d'un territoire impactent les pratiques spatiales et réciproquement. Lorsque ces propriétés sont nécessaires pour la méthode mobilisée, on tente alors d'utiliser des distances mathématiques plus appropriées comme la distance de Manhattan* de la famille des distances de Minkowski, dans des contextes urbains (Muller 1983). Quels que soient les choix, on aboutit dans ce cas à une matrice de distance D(dij), dite carrée, qui évalue pour l'ensemble des couples d'entités étudiées, leurs éloignements selon le sens choisi. Lorsque l'on travaille sur des entités « zonales », la distance entre entités peut être fondée sur la notion de contiguïté, basée sur l'existence ou non d'une frontière commune entre 2 zones (figure 1.a) et par extension, sur le nombre minimal de frontières à traverser pour aller d'une zone à une autre. D'un point de vue thématique, cette notion a beaucoup de sens, les lignes formant un polygone pouvant être assimilées à une frontière. Contrairement à une distance qui est une mesure, la contiguïté dans sa forme première binaire (contigüe, non contigüe) s'assimile à une relation : i est contigüe à i'.

- Une fois définie la métrique associée à l'espace relativement à la question considérée (distance euclidienne ou distance à vol d'oiseau, distance par les voies de communication…), on peut ensuite concevoir des notions telles que les **voisinages** : pour chaque entité i, on identifie l'ensemble des entités i' qui lui sont « proches » au sens de cette distance. Si R définit la relation de proximité basée sur la métrique choisie, le voisinage d'une entité i s'écrira : V(i)={i', iRi'}. La relation peut être basée sur la proximité territoriale (contiguïté) (figure 1.a), la proximité spatiale ou par réseau, ou encore sur des liens existants (figure 1.b).

[3] Règle qu'il est devenu ensuite classique d'appeler la première loi en géographie.

Les relations spatiales se formalisent comme une liste de couples d'entités en relation.

- Enfin, on peut dépasser l'évaluation binaire d'une relation de voisinage (est ou n'est pas voisin (1/0)) en introduisant plutôt un système de pondération* variant entre 0 et 1 calculé à partir d'une fonction d'**interaction spatiale**. Cette fonction permet d'évaluer pour tous les couples de points (i,i') leur potentialité d'interaction compte-tenu de leur éloignement. La fonction répond à des hypothèses de décroissance du potentiel d'interaction entre deux entités avec l'augmentation de la distance qui les sépare, en référence à la loi énoncée par Tobler. Selon sa forme analytique on peut l'interpréter comme une intensité ou encore une probabilité d'interaction compte tenu de la distance (figure 1.c). On produit ainsi une matrice de pondération appelée matrice de poids spatiaux (spatial weights matrix) qui sera notée ici W (wii') où wii'= f(dij) (figure1.c).

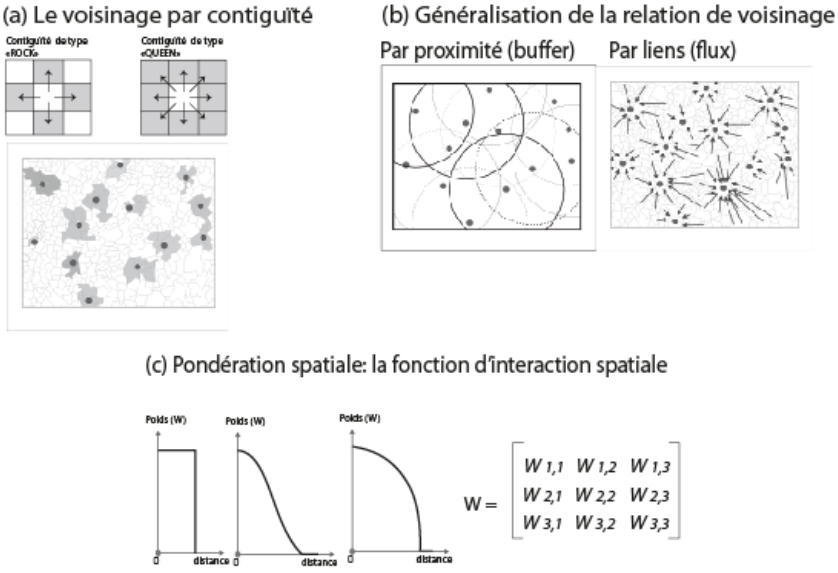


Fig. 1 : Formaliser les relations dans l'espace

Les formalismes de l'espace que nous venons de présenter, permettent de rendre compte de la structure d'espacement et des potentialités d'interaction entre les lieux. L'ensemble de ces éléments formels seront mobilisés au cours des développements suivants. Cette introduction nous parait évidemment nécessaire à la compréhension des méthodes, mais permet aussi de donner du sens à leurs usages dès lors qu'ils permettent d'intégrer les éléments de connaissances des pratiques spatiales associés au phénomène analysé. Les méthodes présentées ci-après peuvent être appréhendées comme une boîte à outils à utiliser en faisant varier les échelles et en multipliant les points de vue pour enrichir la connaissance.

3 Maillages territoriaux et analyse de l'autocorrélation spatiale

De nombreux processus en géographie conduisent à des organisations spatiales dites « spatialement autocorrélées » : ces organisations présentent des motifs illustrant la dépendance spatiale entre les valeurs. C'est le cas des structures spatiales en gradient, ou en régionalisation. L'**autocorrélation spatiale** est à la fois le phénomène de dépendance entre les valeurs en fonction de la distance géographique[4] et la mesure qui est faite de cette caractéristique*. Nous présentons ici des méthodes permettant de la mesurer, tout d'abord globalement sur l'espace étudié, puis localement.

3.1 Approche globale

Les processus qui conduisent à des structures spatiales autocorrélées sont divers. On peut mesurer le degré d'autocorrélation spatiale d'une configuration sans connaître le processus qui y a conduit. Ce peut être une autocorrélation positive où plus les valeurs sont proches géographiquement, plus elles sont proches en valeurs (figure 2.a) ou une autocorrélation négative où à l'inverse plus les valeurs sont proches plus elles sont différentes (figure 2.c). En cas d'absence d'autocorrélation l'organisation peut présenter une alternance des 2 situations précédentes (figure 2b).

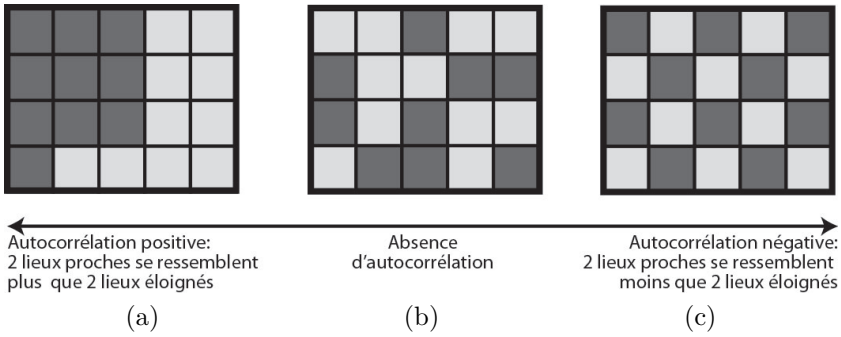


Fig. 2 : Autocorrélation spatiale et organisation spatiale (d'après Pumain, Saint-Julien 1997)

Mesurer l'autocorrélation spatiale revient à mesurer l'intensité de la relation entre « la proximité des lieux » et leur « degré de ressemblance » au regard de cette variable. Pratiquement, la mesure de l'autocorrélation spatiale s'effectue à partir de données zonales. Les indicateurs d'autocorrélation spatiale rapportent une mesure de l'hétérogénéité locale pour une variable donnée, avec la même mesure au niveau global.

L'hétérogénéité locale se mesure par exemple avec la variance* locale, c'est-à-dire une variance calculée sur un ensemble de couples de lieux considérés comme « voisins » au sens du phénomène étudié.

Pour une variable Z, la variance totale (ou globale) s'écrit :

[4] Deux mesures faites dans l'espace ont plus de chance d'être proches en valeurs qu'elles sont proches géographiquement (*cf.* 2.2).

$$\mathrm{var}(z) = \frac{\sum_i (z_i - \bar{z})^2}{n} = \frac{1}{n * (n-1)} \sum_i \sum_{i'} (z_i - z_{i\prime})^2$$

(1)

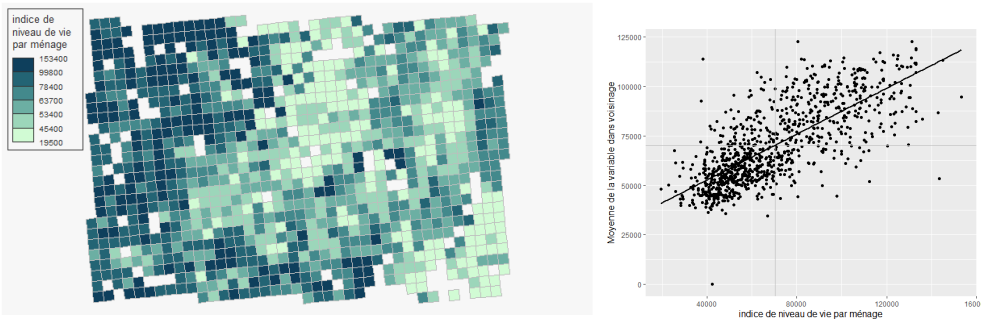
Reprenant la deuxième écriture de la variance, la variance locale, par analogie s'écrira :

$$\mathrm{var}_L(z) = \frac{1}{2K} \sum_i \sum_{i\prime \in V(i)} (z_i - z_{i\prime})^2$$

(2)

Où *V(i)* désigne l'ensemble des entités i', voisines de i et K le nombre de couples *(i,i')* d'entités voisines.

La variance locale regarde l'hétérogénéité moyenne de ce qui se passe en *un lieu* et ce qui se passe *autour*. Une approche graphique de la notion d'autocorrélation spatiale peut être faite en traçant le nuage des points (Z_i, $Z_{V(i)}$). Si celui-ci présente une forme linéaire alors c'est que l'organisation spatiale présente des signes de variations liées aux espacements et donc à l'espace (figure 3b).



(a) (b)

FIG. 3 : Distribution de l'indice de niveau de vie par maille de 200 m (source INSEE) – (a) Configuration spatiale – (b) graphique de Moran croisant pour la valeur en chaque maille avec la valeur moyenne* de son voisinage.

La formalisation de la variance locale présentée en (2) est définie relativement à une relation de voisinage fondée sur la contiguïté entre les lieux, relation de forme binaire : *i'* est ou, n'est pas dans le voisinage de *i*. Cette écriture se généralise (2bis) en introduisant un système de pondération (***Wii'***) comme présenté en 2.3, qui permet ainsi de généraliser la notion de voisinage à des appartenances floues et des fonctionnements dans l'espace. Ainsi (***Wii'***) désigne :

- soit toute forme de voisinage, avec $w_{ii'} = 1$ quand i et i' sont des zones voisines et $w_{ii'}$=0 sinon

- soit une pondération non binaire liée à une proximité ou une interaction spatiale $f(d_{ii'})$. Dans ce dernier cas, tous les couples de valeurs contribuent au calcul mais avec des poids inversement proportionnels à la distance géographique qui les sépare.

$$\mathrm{var}_L(z) = \frac{1}{\sum_{ii\prime} w_{ii\prime}} \sum_i \sum_{i\prime} w_{ii\prime} \left(z_i - z_{i\prime}\right)^2$$

(2bis)

Ce type de formalisation peut s'adapter à d'autres statistiques, comme la covariance* par exemple. C'est cette logique qui sous-tend les **indices**[5] d'autocorrélation spatiale que nous présentons ci-après.

Les deux indices d'autocorrélation les plus utilisés sont :

- ***L'indice de Geary*** est fondé sur la variance locale

$$C = \frac{(n-1)\sum_i \sum_{i\prime} w_{ii,} \left(Z_i - Z_{i\prime}\right)^2}{2\left(\sum_i \sum_{i\prime} w_{ii,\prime}\right) \sum_i \left(z_i - \bar{Z}\right)^2}$$

(3)

Avec n le nombre d'individus, Zi et Zi'les valeurs de la variable Z en i et i '
$\bar{Z}$ la moyenne de la variable Z
Et W_{ii} la pondération spatiale du couple (i,i')

Si l'indice de Geary, noté C (3), est simple à comprendre dans sa formalisation, son usage est troublant car il ne s'interprète pas comme un coefficient de corrélation. Ses valeurs sont toujours supérieures à 0 ; une valeur de 1 correspondant à une absence d'autocorrélation, et il varie à l'inverse de ce qu'il mesure : une valeur tendant vers 0 correspondant à une autocorrélation forte et positive, alors que plus la valeur s'approche de 2 et plus la configuration est autocorrélée fortement mais négativement. Afin d'avoir une interprétation plus simple, on travaille généralement avec l'indice construit à partir de cette forme initiale, à savoir 1-C, qui vaut 0 lorsque l'autocorrélation est nulle, et dont le signe varie ensuite en fonction d'une autocorrélation positive ou négative.

[5] Un indice est un nombre sans unité spécifique qui permet de faire des comparaisons de plusieurs valeurs par rapport à cet indice. Un indice est souvent utilisé pour faire des comparaisons dans le temps et dans l'espace. Les indices, dits élémentaires, résultent d'un calcul simple comme le ratio entre les valeurs d'une variable : par exemple l'évolution d'une variable X entre une date 0 et t se traduira par un indice I dont la valeur sera défini par le rapport entre X_t/X_0. D'autres indices, dits synthétiques, résultent d'un calcul plus compliqué qui combine les valeurs de plusieurs mesures ou variables comme c'est le cas ici pour les indices de Geary et Moran.

- ***L'indice de Moran*** est fondé sur la covariance locale.

$$I = \frac{n}{\sum_i (z_i - \bar{z})^2} * \frac{\sum_i \sum_{i,} w_{ii\prime} (z_i - \bar{z}) (z_{i\prime} - \bar{z})^2}{\left(\sum_i \sum_{i,} w_{ii'}\right)}$$

(4)

Avec n le nombre d'individus, Zi et Zi' les valeurs de la variable Z en i et i '
$\bar{Z}$ la moyenne de la variable Z
Et W_{ii} 'est la pondération spatiale du couple $\left(i, i'\right)$

L'indice de Moran s'interprète comme le rapport entre la covariance locale (2ème terme) et la variance totale (dénominateur du 1er terme). Cet indice est préféré à celui de Geary car il varie et s'interprète comme un coefficient de corrélation* entre -1 et +1.

À l'instar d'un coefficient de corrélation il est possible de tester la significativité de ces coefficients. Par exemple, pour l'indice de Moran, on peut construit la statistique :

$$\frac{I - E(I)}{\sqrt{\text{var}(I)}} \sim N(0,1)$$

Où *E(I)* est l'espérance de I qui tend vers 0 quand le nombre d'entités n augmente. Nous ne détaillons pas ici le calcul de l'écart-type* *S(E(I))* qui est une fonction du système de pondération W qui décrit la relation de voisinage choisie (Cliff et Ord, 1973).

À titre d'exemple, le coefficient de Moran associé à la configuration spatiale présentée dans la figure 3, est de 0.58 et est très significatif (p-value $< 2.2^{e}$-16)

3.2 Approche locale

L'autocorrélation spatiale est une mesure globale, elle caractérise l'agencement dans son ensemble, c'est-à-dire la ressemblance moyenne entre des entités proches. Mais on peut observer aussi une certaine hétérogénéité spatiale dans cet agencement, c'est-à-dire que certaines zones peuvent être plus caractéristiques d'une forme d'agencement que d'autres. À la suite de l'indice de Moran, anselin (1995) propose une version locale de cet indice : LISA (*Local Indicator of Spatial Association).* Cet indice donne pour chaque entité, sa contribution au calcul de l'indice global. En effet, l'indice global est une somme de contributions de la variabilité des voisinages de chaque entité *i*.

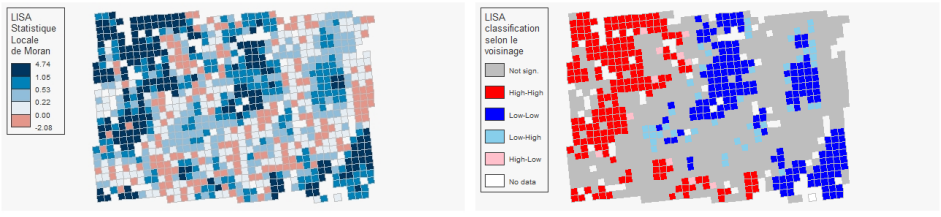
Chaque contribution est notée I_i. Et s'écrit :

$$I_i = \frac{(z_i - \bar{Z}) \sum_{i\prime} w_{ii} (z_{i\prime} - \bar{Z})}{\text{var}(Z)}$$

Avec

$$I = \frac{1}{\left(\sum_i \sum_{i,} w_{ii}\right)} * \sum_i I_i$$

Cet indice peut être cartographié (figure 4a), et permet de différencier les situations localement. Il est ainsi possible d'identifier les situations localement homogènes (fortes valeurs de I_i) caractérisant des concentrations locales de fortes ou faibles valeurs, et les situations localement hétérogènes (faibles valeurs de I_i) soit parce que l'on observe de fortes valeurs de Z en i et de faibles valeurs dans son voisinage, ou réciproquement. Un deuxième traitement permet d'isoler ses situations, en ne « coloriant » que les mailles dont les valeurs d'autocorrélation locale, I_i, contribue significativement à la valeur de l'autocorrélation globale. On fait ainsi apparaître des régions (clusters). Parmi elles on différencie quatre types (figure 4b) : celles qui contribuent à une autocorrélation spatiale positive, soit par de fortes valeurs dans un voisinage de fortes valeurs (high-high) ou de faibles valeurs dans un voisinage de faibles valeurs (Low-Low) ; et celles qui contribuent à une autocorrélation négative soit par fortes valeurs dans un environnement de valeurs plus faibles (High-Low, soit par de faibles valeurs dans un environnement de fortes valeurs (Low-High) (Anselin 1995).



(a) (b)

FIG. 4 : a) Carte des Indices Locaux d'autocorrélations spatiales. Plus l'indice est fort, plus une maille ressemble à son voisinage. b) Cartographie des classes de ressemblances significatives entre une maille et son voisinage, identifiées par le LISA.

4 Semis de points et analyse du plus proche voisin

Les méthodes d'analyse des semis de points ont pour objectif de décrire la forme d'organisation spatiale d'un ensemble de localisations. Les deux méthodes les plus connues et les plus utilisées sont la méthode des quadrats et la méthode des plus proches voisins. La première, la méthode des quadrats, procède à un calcul *via* un maillage régulier et repose sur un comptage du nombre de points par maille. La

seconde s'appuie sur la notion *de plus proche voisin*. Nous ne présentons ici que cette dernière qui permet d'une part des développements dans un cadre descriptif ou exploratoire (4.1), mais aussi confirmatoire (4.2) et sur laquelle s'appuient des prolongements (4.3). Pour la méthode des quadrats nous renvoyons le lecteur à la bibliographie (Jayet 1993, Pumain, Saint-Julien 1997).

4.1 Espacements, notion de « plus proche voisin » et mesures associées

Lorsque l'on dispose d'un semis de points, la description de l'organisation du semis se fonde sur l'évaluation des espacements entre les points, résumés à l'ensemble du semis.

Une fois choisie une mesure de distance (au sens du 2.2) on s'intéresse à une distance spécifique pour chaque localisation : la « distance au plus proche voisin ». Cette distance peut être mesurée pour le voisin le plus proche. Elle est aussi généralisée par une moyenne des distances aux 3, ou 5 ou n voisins les plus proches[6] (fig. 5), permettant de différencier les cas de voisinage dense et moins dense. Cette mesure sera utilisée dans un cadre inférentiel, mais il est intéressant d'utiliser la variable ainsi construite dans un cadre descriptif, pour rendre compte de l'hétérogénéité des espacements dans l'espace ou pour être intégrée dans d'autres analyses, comme variable explicative dans une démarche confirmatoire (une régression par exemple).

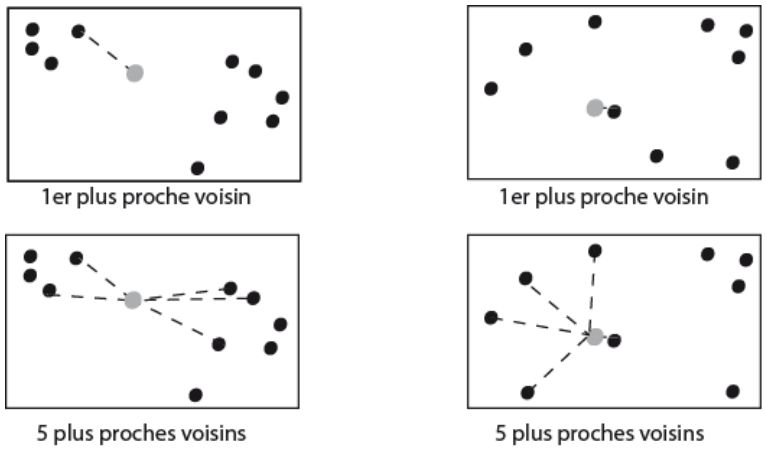


Fig. 5 : La notion de « plus proche(s) voisin(s) ».

[6] Approche que l'on retrouve dans le chapitre 224

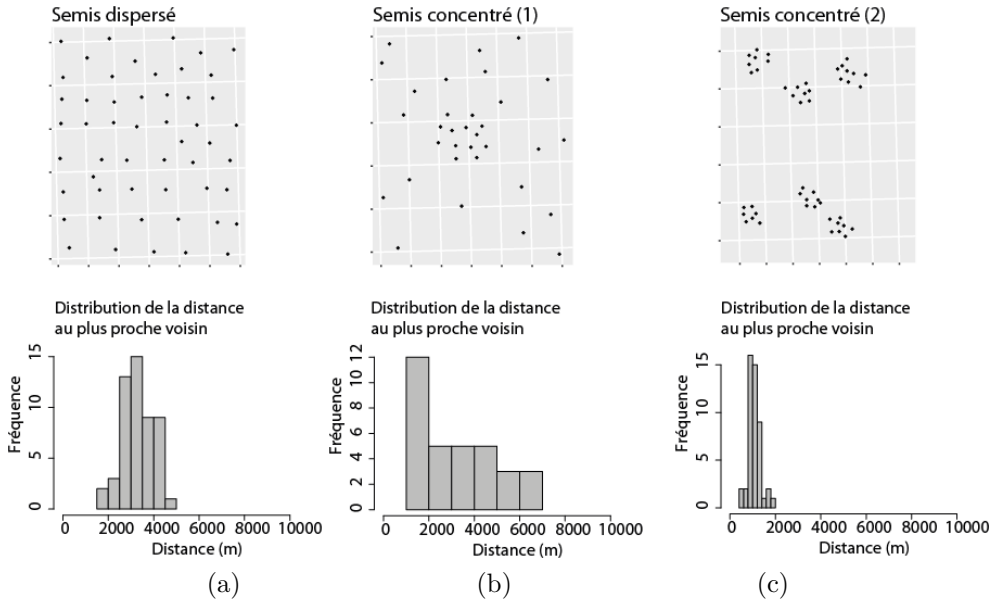


FIG. 6 : Types de semis et distributions associées de la variable « distance au plus proche voisin ».

Dans une approche descriptive, l'analyse de la distribution la variable « distance au plus proche voisin » permet de comprendre l'organisation du semis, de comparer ces distributions entre plusieurs semis (dates, régions…). Elle permet d'approcher les caractéristiques de l'organisation globale. Plus la distribution est asymétrique, plus on va pouvoir identifier soit une structure concentrée (sur-représentation des petites distances), soit une structure régulière (sur-représentations des grandes distances). La figure 6 illustre ces différentes distributions pour trois types de semis. En colonne (a) on trouve l'histogramme* associé à une distribution spatiale dispersée, montrant une concentration forte des valeurs entre 2000 et 4000 m. Les colonnes (b) et (c) illustrent deux cas différents de semis concentrés. Dans le cas de la colonne (c), le semis est très concentré, par paquets homogènes, et cette concentration se traduit par une concentration des valeurs de la « distance au plus proche voisin » entre 0 et 2000 m. Alors que le semis de la colonne (b) est irrégulièrement concentré, et l'histogramme des distances au plus proche voisin reflètent à la fois la concentration en paquet avec une concentration de valeurs entre 1000 et 2000m, mais aussi une hétérogénéité avec des cas où cette valeur peut atteindre 7000m. Ce type d'approche descriptive sous-tend une partie de l'analyse du 2.2.4.

4.2 Analyse d'un semis de points : la méthode du « plus proche voisin »

La méthode dite « du plus proche voisin » inscrit dans un cadre inférentiel la description d'un semis de points par le calcul de la moyenne de la variable « distance au plus proche voisin » (d_{ppv}). La méthode compare cette valeur à la distance moyenne que l'on aurait obtenue dans le cadre d'une expérience aléatoire donnant un semis dit

« aléatoire » (*cf.* 2.3). Il s'agit de l'espérance de la distance moyenne d'un point à son plus proche voisin ($d_{\text{théo}}$) dans le cas où le processus aléatoire est un processus de Poisson[7] (les coordonnées des points sont issues d'un tirage aléatoire)

$$d_{\text{théo}} = \frac{0.5}{\sqrt{\frac{n}{S}}}$$

Où S est la superficie de l'espace étudié et n le nombre de points localisés sur cet espace.

Un premier indicateur résumant la forme globale du semis est donné par le rapport entre la moyenne observée :

$$\overline{d_{ppv}}$$

Et la valeur théorique de cette distance moyenne :

$$d_{\text{théo}}$$

Soit :

$$R = \frac{\overline{d_{ppv}}}{d_{\text{théo}}}$$

La comparaison des deux valeurs permet d'identifier la tendance de l'organisation du semis : concentré, régulier, aléatoire. La figure 7 illustre différents cas typiques de semis et de valeurs associées du rapport R : 0 pour un semis concentré ; 2.15 pour un semis espacé c'est-à-dire régulier ; le rapport vaut 1 lorsque le semis est « aléatoire ».

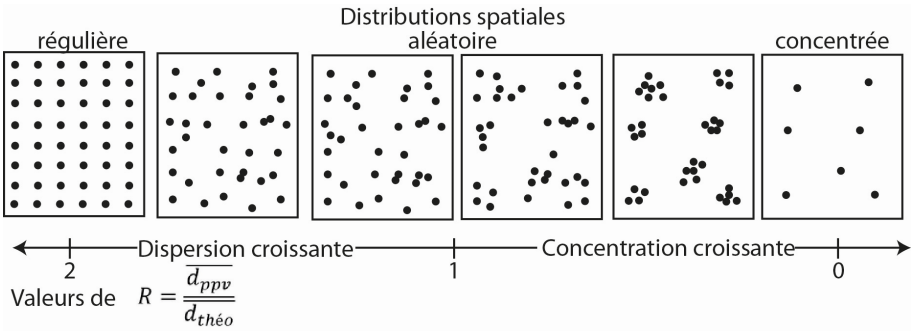


FIG. 7 : Caractérisations de différents types d'organisation d'un semis de points.

La statistique basée sur l'écart relatif entre la moyenne théorique et la moyenne observée permet de tester l'hypothèse selon laquelle le semis observé peut être ou non issu d'un processus aléatoire tel que le processus de poisson.

$$\frac{|d_{ppv} - d_{\text{théo}}|}{\sqrt{\text{var}\,(d_{\text{théo}})}} \sim N(0,1) \quad \text{avec} \quad \text{var}\,(d_{\text{théo}}) = \frac{0.261}{\sqrt{\frac{n^2}{S}}}$$

[7] Sur la loi de Poisson *cf.* 7.

La valeur de la statistique obtenue permet de conclure si l'arrangement des points observés diffère significativement ou non d'un arrangement produit par un processus aléatoire, et si c'est le cas, le rapport R permet d'identifier la tendance du semis entre concentration et dispersion.

4.3 Décrire les organisations multiscalaires

La méthode du plus proche voisin donne une mesure globale de l'organisation, mais elle est fortement dépendante de la taille de la zone d'analyse et de sa localisation. En outre, lorsque les processus sous-jacents sont multiples et de portées géographiques variables, l'organisation spatiale du semis analysé peut être plus complexe et présenter des organisations multiscalaires dont la description nécessite de dépasser les seuls qualificatifs de « semis concentré » ou « semis dispersé » (Figure 8).

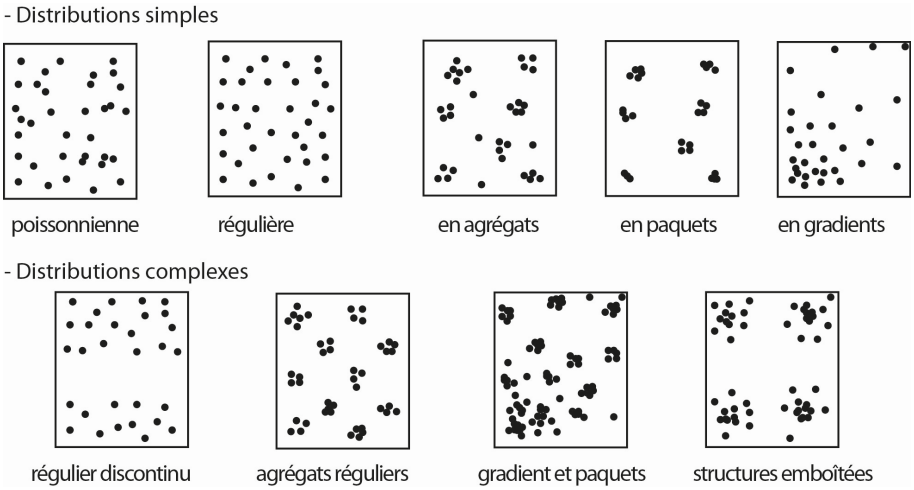


FIG. 8 : Quelques motifs types de dispersion spatiale identifiés en écologie (d'après Chessel 1978)

Plusieurs prolongements ont été proposés. Nous illustrons deux d'entre eux, les plus utilisés.

Le premier appelé « le plus proche voisin raffiné » consiste à construire la courbe cumulée de la variable « plus proche voisin » et de construire ainsi la fonction $\boldsymbol{G(r)}$ qui évalue la part des points dont le plus proche voisin est à l'intérieur d'un cercle de rayon $\boldsymbol{r}$:

$$G(r) = \frac{\#\{i/d_{ppv}(i) < r\}}{n}$$

où # désigne la fonction « cardinal » c'est à dire la fonction qui compte le nombre d'éléments d'un ensemble et n le nombre total d'entités

La forme et en particulier la pente de la fonction $\boldsymbol{G(r)}$ quand $\boldsymbol{r}$ varie, permet de rendre compte de la distribution* des espacements et ce à plusieurs échelles $\boldsymbol{(r)}$. Plus

les points sont concentrés, plus la pente de la courbe au démarrage est forte, traduisant le fait qu'un grand nombre d'entités ont des « plus proches voisins » à faible distance. Lorsque la courbe comporte des paliers, c'est que le semis comprend des emboîtements de structures.

Le second développement que nous citons et illustrons est celui de la fonction K (Ripley, 1976 et 1981) qui s'appuie sur la même logique de dénombrement mais pour toutes paires de points dont la distance est inférieure à ***r*** (figure 9).

$$K(r) = \frac{1}{n} \sum_{i=1}^{n} \frac{N_i(r)}{\frac{n}{S}}$$

où $N_i(r)$ désigne le nombre de points dans un cercle de rayon r autour de i
S est la superficie de l'espace d'étude et n le nombre total de points.

Ainsi à toute distance ***r***, on associe le nombre total de paires de points espacés d'une distance inférieure à ***r***. Cette fonction retranscrit mieux la variabilité globale des espacements, notamment dans le cas d'agrégats, et les effets de seuils rendent compte des différentes tailles de ces agrégats (Pélissier et Goreaud 2001, Bevan et Conolly, 2006). La figure 9 illustre la fonction pour les 3 cas présentés en figure 6, ci-dessus. La courbe noire sert de référentiel : c'est la courbe théoriques associée à un semis aléatoire. La courbe associée au semis dispersé (courbe « rouge ») présente des densités* de voisinage très faibles localement, et rejoint une densité « aléatoire » quand le rayon atteint 4500m. Tandis que les courbes « bleues » associées aux semis concentrés, présentent de fortes densités de voisinages de faible portée. Pour le semis le plus concentré (courbe « bleue claire ») on voit nettement apparaître la taille des agglomérats autour de 4000m lorsque la pente de la courbe s'infléchit.

5 Du discret au continu

Dans le cadre de ce chapitre, les observations sont discrètes (ponctuelles ou zonales). Dans la pratique, on rencontre souvent le besoin de produire une représentation continue du phénomène observé (*cf.* 2.2.6). C'est une nécessité si l'on dispose de points échantillonnés d'un champ continu et dans ce cas on utilise des méthodes d'*interpolations spatiales* dont les *géostatistiques* font partie (5.1). Il est d'usage de le faire aussi lorsque les points correspondent à des événements (points avec une seule valeur existe/ n'existe pas) et que l'on cherche à rendre compte de la *densité des événements*, ce qui conduit à construire une surface de densité. Dans ce cas les méthodes mobilisées sont des *méthodes d'estimation par noyau* (kernel density estimation) (5.2). Enfin on peut aussi avoir recours à cet objectif lorsque les points représentent des localisations d'objets plus complexes (bâtiments, ville, centroïdes d'une maille administrative…) qui structurent des territoires et à partir desquels s'organisent des mouvements. Une surface continue, construite sur la base de fonctions d'interactions spatiales, peut être un moyen de rendre compte de fonctionnements dans l'espace. Dans ces cas on parlera de *méthodes de lissage* (5.3).

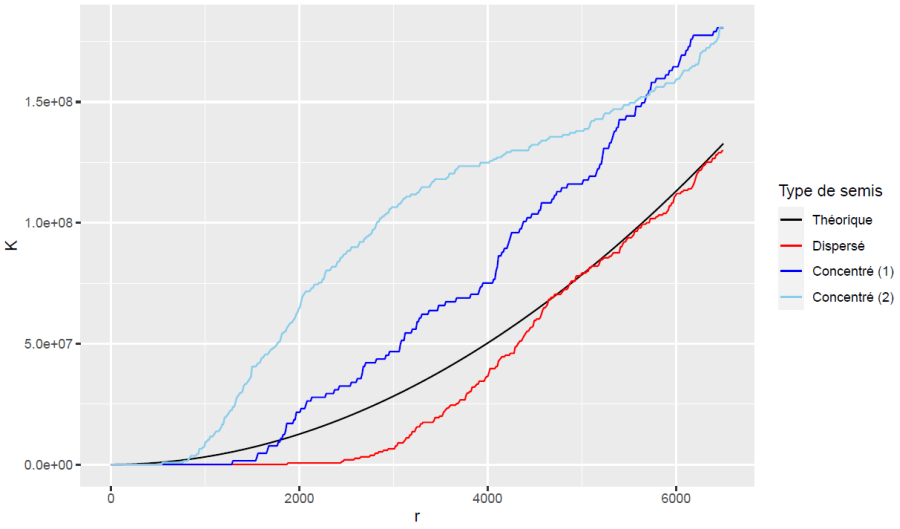


FIG. 9 : La fonction K pour détecter les formes des semis et les espacements des agrégats.

Nous présenter brièvement les fondements de chacune des familles, en les mettant en regard sur la base des éléments formels précédemment présentés. Nous renvoyons à la lecture de différentes références très complètes pour approfondir ces approches.

La finalité de l'ensemble des méthodes qui sont présentées dans cette partie, et qui permettent de passer de données discrètes à des représentations continues de l'espace, s'énonce donc de manière générique de la façon suivante :

Connaissant les valeurs Z associées à n localisations $(z_1, z_2, \ldots z_n)$, on cherche $\hat{z}_i$ en tous points de l'espace.

5.1 Méthodes d'interpolations spatiales :

L'**interpolation spatiale** consiste, à partir de valeurs échantillonnées dans l'espace pour un phénomène donné (z_i), d'estimer la valeur de ce phénomène pour tous points de l'espace $(z_{i'})$.

Les méthodes d'interpolation sont nombreuses et sont généralement classées en deux familles : les méthodes déterministes et les méthodes stochastiques. Dans le premier cas les valeurs sont calculées à partir des valeurs observées sur la base de fonctions mathématiques, alors que dans le second cas, les valeurs sont calculées à partir des méthodes statistiques d'estimation tenant compte de la variabilité du phénomène échantillonné dans l'espace. Nous présentons rapidement les logiques associées à chacun des cas à partir d'un exemple de méthode à savoir l'inverse à la distance IDW (inverse distance weighting) pour les méthodes déterministes et le Krigeage* pour les méthodes stochastiques.

1- L'IDW est une méthode fondée sur l'hypothèse qu'en tous points de l'espace, les valeurs proches se ressemblent plus que les valeurs éloignées. Ainsi la valeur calculée

en un point quelconque de l'espace $z_{i'}$, dépend plus d'une valeur observée z_i proche que d'une valeur observée z_i éloignée. On retrouve ici le système de pondération $w_{ii'}$ déjà évoqué ci-dessus (2.3) modulant l'influence de la valeur de i sur i' en fonction de leur éloignement.

$$\hat{z}_{i'} = \sum_i w_{i,i'} z_i$$

Où w_{ij} est le poids associé au couple *(i,i')*.

Cette pondération est une fonction inverse de la distance, qui définit une fonction d'interaction spatiale :

$w_{ii'} = f(d_{ii'})$.

La fonction de la distance la plus répandue et utilisée est sans aucun doute

$$f\left(d_{i,i\prime}\right) = \frac{1}{{d_{i,i\prime}}^2}$$

2- Le Krigeage est certainement la méthode d'interpolation spatiale la plus répandue dans les domaines étudiant les phénomènes continus dans l'espace. C'est une méthode d'interpolation reposant sur une approche statistique de la variabilité des valeurs mesurées dans l'espace. Méthode d'estimation complexe, nous en donnons ici les grandes lignes. Cette méthode comporte plusieurs formes. Nous décrivons rapidement les grandes lignes associées au Krigeage ordinaire.

À partir de valeurs observées localisées (figure 10.a), la méthode d'estimation des valeurs dans tout l'espace repose sur une première modélisation, celle de la variabilité des valeurs observées dans l'espace : c'est le variogramme*[8]. À l'instar d'une variance locale, le variogramme de la variable ***Z*** décrit les variations moyennes de valeurs observées entre deux points en fonction de leur espacement ***h*** (figure 10.b). Il s'écrit :

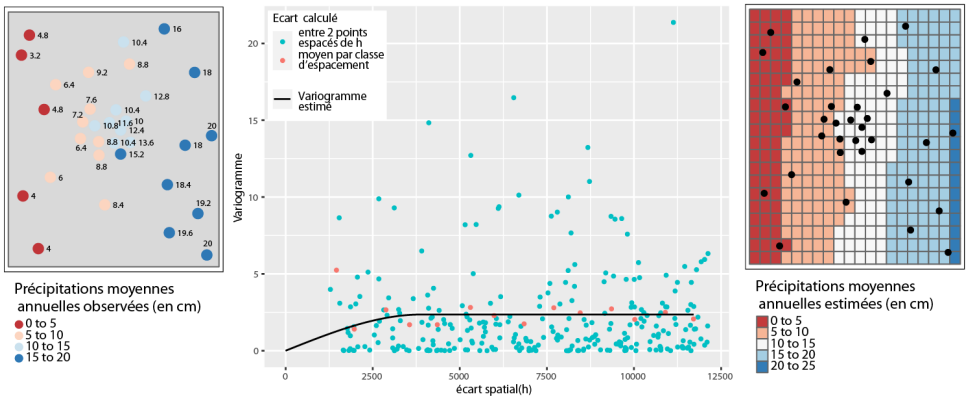
$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(i) - Z(i+h)]^2$$

Où ***Z(i)*** désigne la valeur de ***Z*** pour un point ***i*** et ***Z(i+h)*** la valeur de ***Z*** pour un point situé à distance ***h*** de ***i***

N(h) désigne le nombre de couples de points ***(i, i+h)*** distants de ***h***

L'hypothèse associée à cette première étape de modélisation est la non-stationnarité* des variations de valeurs, c'est-à-dire qu'en moyenne la différence de valeurs entre 2 points dépend de leur espacement. À partir de ce nuage de points, on regroupe les points par pas de distance (classe* de distance), et on identifie la fonction qui ajuste

[8] On parle aussi de semi-variogramme, la formule du variogramme comprenant une division par 2. Nous utiliserons ici le terme de variogramme pour désigner cette quantité.



(a) (b) (c)

FIG. 10 : Calcul de variogramme et estimation par méthode du Krigeage
(a) Semis de points observés (b) variogramme expérimental et ajusté (c) Surface estimée par la fonction de variogramme ajustée

au mieux le nuage des points moyens ainsi produit (variogramme échantillonné). Cette fonction décrit la variabilité moyenne des valeurs de $\boldsymbol{Z}$ en fonction des espacements (figure 10.b). Si l'on observe que cette variabilité n'est pas homogène selon différentes directions dans l'espace, c'est-à-dire que le phénomène est anisotrope, il est possible de décomposer le variogramme selon plusieurs directions (par exemple Nord-Sud, Est-Ouest, ou encore Nord-Est/Sud-ouest…) (*cf.* chapitre 2.2.6). Le choix du modèle* du variogramme contient une part de subjectivité. Plusieurs modèles peuvent être candidats. Les points importants sont le comportement à l'origine et les paliers. Le choix s'effectue le plus souvent sur la base d'autres informations ; de la connaissance du terrain et de l'expérience.

Le modèle permet ensuite de calculer une estimation des valeurs de $\boldsymbol{Z}$ en chaque point d'une grille de la taille de la première valeur de l'écart $\boldsymbol{h}$ (figure 10.c) L'estimation est une somme pondérée des valeurs observées, où le système de pondération optimal est le résultat d'une résolution d'un système d'équations linéaire à partir du variogramme estimé (Kriging simple). L'intérêt du Krigeage est qu'il produit non seulement un ensemble de valeurs estimées sur une grille régulière, mais comme toutes les méthodes statistiques il complète ces estimations par un intervalle de confiance*, ce qui permet de relativiser la surface ainsi produite (Arnaud et Emery, 2000). Le chapitre 2.2.6 illustre une application de cette méthode, et la figure 8 de ce chapitre montre la carte des erreurs estimées par le modèle

5.2 Méthodes d'estimation des densités par les noyaux

Ces méthodes concernent les semis de points de type « événements » (crime, maladie, …), pour lesquels on cherche à illustrer l'intensité du phénomène en représentant la densité de probabilité* du phénomène sur une surface continue. Un calcul de la densité globale des points sur la superficie de la région étudiée, est une forme de gé-

néralisation qui repose sur l'hypothèse que le phénomène peut être observé en tous points de l'espace, les événements observés sont alors considérés comme des échantillons. La méthode d'estimation par noyaux (**Kernel Density Estimation KDE**) est ici utilisée en dimension 2. C'est une méthode non-paramétrique, reposant sur une fonction de densité théorique ***K*** (le noyau) et une portée de lissage ***r***. En tous points ***x*** de l'espace on estime la densité de probabilité d'apparition d'un événement par :

$$\delta_K(x) = \frac{1}{r^2} \sum_{i=1}^{n} K\left(\frac{d(x,i)}{r}\right)$$

Où ***k()*** est une fonction de probabilité définie sur [0,1],
r est le rayon du noyau (« bandwidth »)
d(x,i) la distance entre le point ***x*** et un des ***n*** points observé ***i***

La forme la plus simple de cette méthode, est celle où le noyau est une fonction rectangulaire, ce qui peut se représenter par une fenêtre circulaire de rayon ***r*** (figure 11.b).

Dans ce cas, en tous points ***x*** de l'espace, la densité se calcule avec l'équation suivante :

$$\delta_C(x) = \frac{N(C(x,r)}{\pi r^2}$$

Où ***C(x,r)*** est le cercle de centre **x** et de rayon ***r***,
Et ***N()*** est la fonction qui dénombre le nombre d'événements observés dans cette surface.

Cela revient à calculer pour chaque point, une densité « autour », à partir du nombre de points-événements contenus dans le cercle (figure 11a).

Le noyau (**K**) peut ainsi prendre différentes formes en s'appuyant sur des fonctions de densités différentes. La figure 11b illustre les plus utilisées (Rectangulaire, Triangulaire, Biweight, Epanechnikov) et la figure 11c présente la forme spatiale d'un noyau associé à la fonction Biweight. La définition de ***r*** dépend du phénomène observé et nécessite une phase de calibrage. Une portée trop petite dessinera la géographie des points, et une portée trop grande aboutira à une densité très homogène. Des développements viennent affiner encore cette question de portée de noyau en proposant une adaptation des portées en fonction du contexte (Baxter *et al.* 1997, Silverman 1986, Nuninger *et al.* 2016).

Ainsi, quelle que soit la forme du noyau, on peut calculer en tout point une densité d'événements à partir des événements voisins, chacun contribuant à cette densité avec une pondération suivant une fonction inverse de la distance qui le sépare du point de calcul. Par analogie, on peut rapprocher cette formalisation des fonctions d'interactions spatiales présentées précédemment.

5.3 Méthodes de lissage

Les méthodes de lissage se distinguent des méthodes précédentes dans le fait qu'il ne s'agit que d'évaluation et non d'estimation comme dans les deux cas précédents

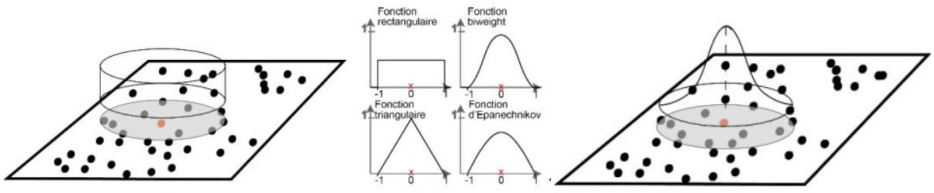


FIG. 11 : Formes des noyaux pour le calcul des densités locales

(Grasland, Mathian, Vincent 2000). C'est donc quasiment plus une question de sémantique que de méthodologie. Elles concernent des données agrégées, localisées dans une maille et pour lesquelles on veut s'abstraire du maillage soit pour des raisons cartographiques, soit pour des raisons thématiques. Dans ce dernier cas, on doit avoir des hypothèses sur les mouvements potentiels de ces effectifs* et qu'il convient de répartir chaque unité de cet effectif sur le territoire environnant.

Les calculs sont identiques aux calculs présentés juste au-dessus pour les densités et reposent sur des fonctions d'interactions spatiales, justifiant qu'une quantité P localisée en i puisse être comptabilisée « autour ». Le « autour » correspond à un voisinage à définir en fonction du phénomène étudié, et le voisinage peut intégrer la notion de « potentiel ». La fonction de potentiel est une fonction d'interaction spatiale.

La valeur en tous points i s'écrit :

$$A_i = \sum_{j=1}^{n} O_j f\left(d_{ij}\right)$$

Où Oj sont les valeurs observées, f(dij) la fonction d'interaction spatiale comme précédemment vue.

La fonction d'interaction spatiale reflète les fonctionnements dans l'espace associé au phénomène analysé. On obtient ainsi des surfaces continues qui peuvent s'interpréter relativement aux mouvements des populations statistiques étudiées ou à l'accessibilité potentielle des ressources (Gauthier *et al.* 2022 part. 2 chap. 4, Nirascou, Lacaze, 2000)

6 Conclusion sur les statistiques spatiales

Les méthodes, dîtes de statistiques spatiales, sont à considérées comme un ensemble formant une boite méthodologique permettant d'explorer et révéler des formes d'organisation des phénomènes étudiés en intégrant la dimension géographique des données. Ces méthodes sont pensées de façon multiscalaire pour identifier des formes ou intégrer des effets de contextes. Le chapitre s'est efforcé de donner un aperçu de la diversité de ces méthodes. Le chapitre renvoie pour chaque famille de méthodes à des références bibliographiques permettant de revenir plus précisément sur des aspects théoriques ou pour approfondir leur opérationnalité.

L'aperçu qui est proposé ici permet de situer les méthodes les unes par rapport aux autres, et aussi relativement à un schéma plus conceptuel d'appréhension de l'espace dans ses différentes granularités, locale ou globale, et dans ses différentes formalisations, territoriale ou spatiale... Cette approche permet de redonner à la dimension spatiale son rôle fondamentale, d'objet d'étude. Les méthodes présentées peuvent alors sortir de leur strict cadre d'usage premier et être utilisées autant en amont d'une démarche, pour créer des données contextuelles (par exemple pour chaque habitat, décrire la dynamique des habitats dans le voisinage) ou estimer des valeurs sur l'ensemble de l'espace et construire des représentations continues d'un phénomène. Ou encore changer de niveau d'observation en construisant des agrégats ou clusters (ou hotspot) ayant un sens pour l'analyse (passer par exemple des habitats aux réseaux d'habitats en faisant l'hypothèse qu'à une certaine proximité ou accessibilité, les habitats fonctionnent ensemble). C'est aussi en replaçant ces méthodes dans une démarche d'analyse que l'on s'aperçoit de la richesse que cet éclairage apporte. Le chapitre renvoie aussi à des exemples de traitement dans la partie 2.2. Beaucoup de développements ont aussi intégré la dimension temporelle, les 2 dimensions, spatiale et temporelle, étant liées et particulièrement mobilisées dans le domaine de l'archéologie. Ces développements permettent dans une approche exploratoire d'approcher les questions de processus.