Analyse de données spatiale :
Problèmes et méthodes d'échantillonnage des données spatiales

I. Introduction

L'échantillonnage spatial est particulièrement important pour la statistique environnementale.

➢ ***Donnée spatiale :***

Toute observation dont on connait non seulement la valeur, mais aussi la localisation.

⇒ Les coordonnées spatiales des objets à traiter constituent une information potentiellement riche pour l'analyse

➢ ***Trois types de données spatiales :***

- *Données spatiales continues (géostatistiques)* : ***Il existe une valeur pour la variable d'intérêt en tout point du territoire***, mais elle ***est mesurée uniquement en un nombre discret de points***.
- *Données spatiales ponctuelles :* Ces données se caractérisent par leur ***distribution dans l'espace***, et pas par la valeur associée à chaque observation. La localisation est la variable aléatoire qu'on étudie.
- *Données spatiales surfaciques :* La localisation des observations est considérée comme fixe, c'est ***leur valeur qui est modélisée selon un processus aléatoire***. Peuvent être agrégées sur une partition du territoire ou réparties en des points précis.

➢ ***Propriétés particulières des données spatiales :***

- Dépendance spatiale : la valeur de l'observation ***i*** influence la valeur de l'observation ***j*** voisine ;
- Hétérogénéité spatiale : l'influence des variables explicatives sur la variable dépendante dépend de la localisation dans l'espace.

➢ ***Analyse spatiale:***

- ***Décrit la structure*** spatiale des observations.
- Quantifie ***la force des interactions*** spatiales.
- Explique ***les mécanismes*** sous-jacents.

➢ ***Objectifs de l'analyse spatiale des données continues :***

Prédire la valeur d'une variable en un point où elle n'a pas été échantillonnée, ainsi que la fiabilité de cette prédiction.

Exemples de données :

- Composition chimique du sol ⇒ industrie minière ;
- Qualité de l'eau ou de l'air ⇒ étude de la pollution ;
- Variables météorologiques ⇒ prévisions météo.

Exemple d'étude :

Avec quelle précision prédire la teneur en polluant d'un sol lorsqu'on n'a qu'un échantillon fini de points de mesure ?

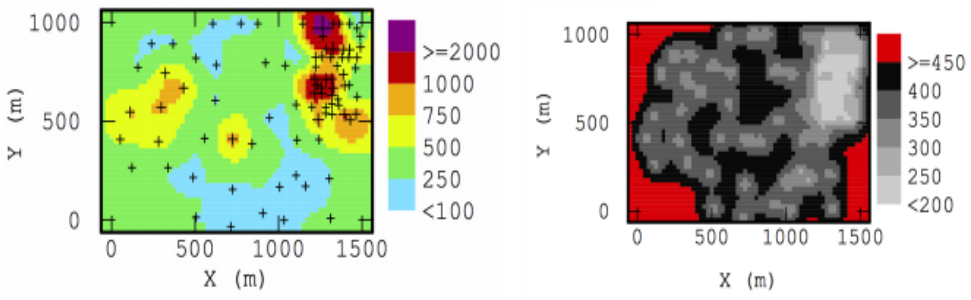


Fig.1. Prédiction de teneur en polluant (mg/kg/m^2) - Ecart type de la prédiction
Source : Manuel GéoSiPol - Mines de Paris

➢ ***Objectifs de l'analyse spatiale des données ponctuelles :***

Quantifier l'écart entre la distribution spatiale des observations et une distribution complètement aléatoire dans l'espace : ***dans quelle mesure les observations sont-elles regroupées dans l'espace ?***

Exemples de données:

- Ecologie ⇒ distribution spatiale de deux espèces d'arbres ;
- Epidémiologie ⇒ distribution spatiale des cas d'apparition d'un virus.

Les trois configurations classiques de points :

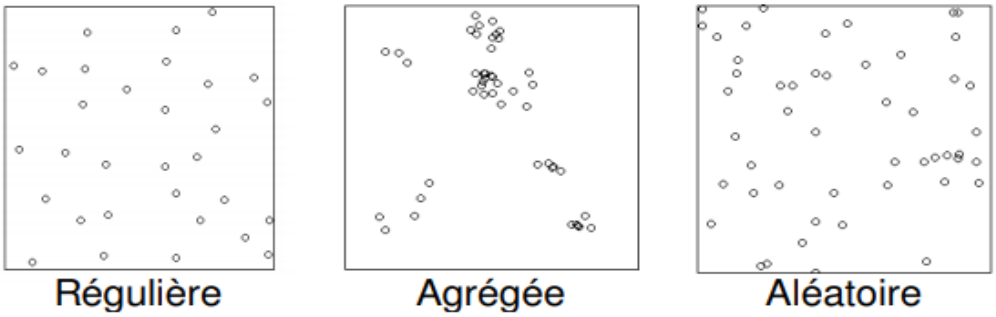


Fig.2. *Les trois configurations classiques de points*
Source : Package spatstat, Floch, Marcon, Puech 2017

Exemples d'étude:

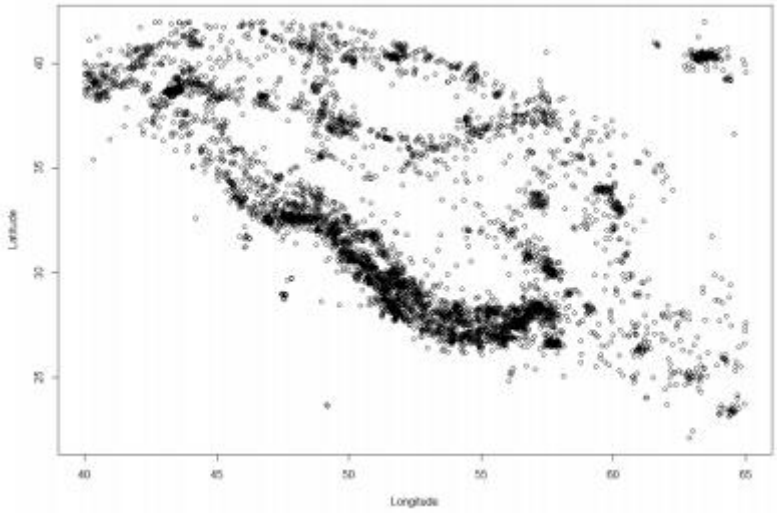


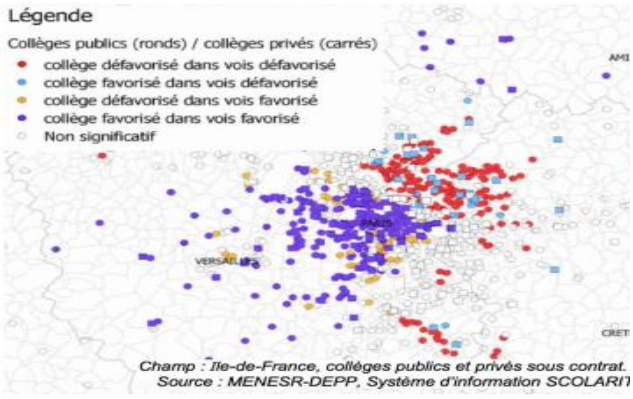
Fig.3. Localisation de 5970 épicentres de séismes en Iran de 1976 à 2016.
Source : Package etas, Floch, Marcon, Puech 2017

➢ ***Objectifs de l'analyse spatiale des données surfaciques :***

- Définir la structure de voisinage des observations.
- Quantifier l'influence qu'exercent les observations sur leurs voisines et sa significativité.

Exemples de données:

- Santé ⇒ taux de malades par région;
- Economie ⇒ croissance du PIB par région.



Exemples d'étude:

Les collèges favorisés sont-ils toujours situés dans un environnement favorisé ?

Source : Givord et al 2015

➢ ***Représentation des données spatiales surfaciques :***

- ***Représentation sous forme d'un polygone :***
  - La région d'étude peut être recouverte exhaustivement par des polygones mutuellement exclusifs.
  - Deux polygones adjacents sont séparés par une frontière commune.
  - Les frontières peuvent naître de discontinuités spatiales : limites administratives, barrières naturelles... Ou être les polygones de Voronoï issus de points particuliers.

Polygone de Voronoï associé au point x_i :

La région de l'espace qui est plus proche de x_i que de tout autre point de l'ensemble d'étude ***S.***

- ***Représentation sous forme d'un point :***
  - Les données "surfaciques" peuvent aussi être des ***points fixes*** du territoire ;
  - Définis par les coordonnées géographiques d'un point particulier : lycée, mairie, point culminant d'une région,...
  - Ou définis géométriquement comme les centroïdes d'un polygone.

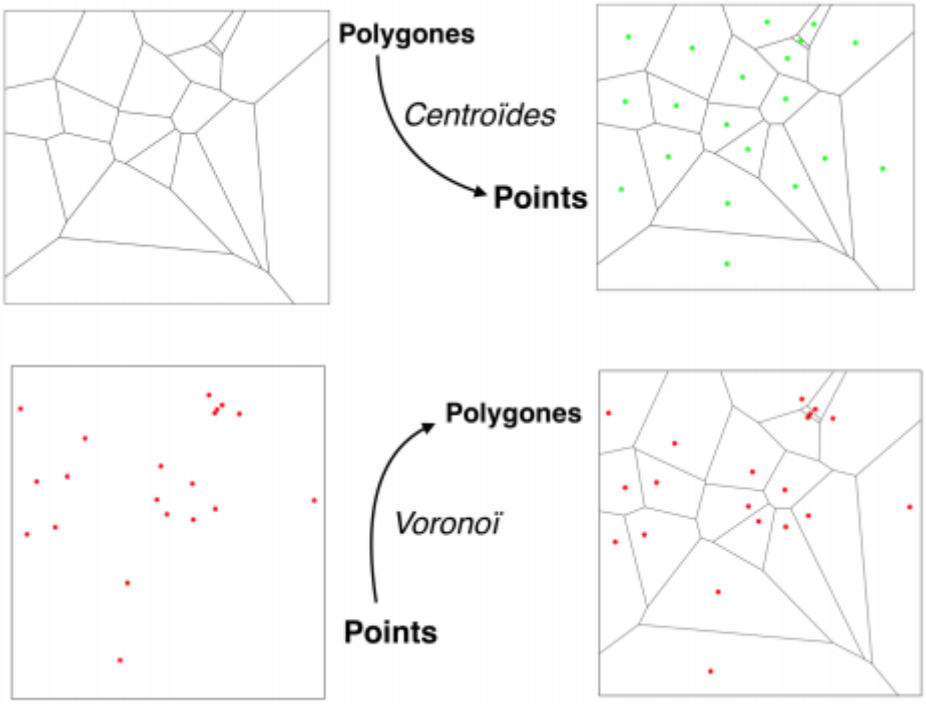
Centroïde c d'une surface S

Minimise la distance quadratique moyenne à tous les points de ***S*** :

$$\min_c \frac{1}{a(S)} \int_S ||x - c||^2 dx$$

a(S) : aire de la surface S

Coordonnées de ***c*** : moyenne des coordonnées de tous les points de ***S***



L'analyse spatiale des données est un domaine ancien et vaste qui regroupe de nombreuses méthodologies, certaines générales et d'autres plus spécifiques à un champ thématique précis. Les logiciels SIG classiques sont donc loin d'être les seuls outils employés en analyse spatiale. Certains viennent de la géostatistique par exemple ou de champs thématiques comme l'écologie du paysage.

Dans le domaine des SIG, on appelle parfois Géotraitements ces fonctions analytiques. Certaines se réalisent en mode vecteur, d'autres en mode raster, d'autres encore fonctionnent dans les deux modes.

II. Problèmes et méthodes d'échantillonnage des données spatiales

II.1. Le problème

Les données spatiales sont souvent autocorrélées. Si dans un échantillon, on sélectionne deux points très proches, on obtiendra des mesures probablement très similaires. On récoltera ainsi moins d'information dans l'échantillon qu'en étalant les points dans l'espace. Quand on sélectionne un échantillon dans l'espace, il est donc intéressant d'étaler les observations. Un ensemble de méthodes ont été développées pour obtenir des points étalés dans l'espace tout en contrôlant les probabilités d'inclusion.

Grafström & Lundström (2013) préconisent même l'utilisation d'un échantillonnage étalé pour des données non-spatiales avec des distances calculées sur des variables auxiliaires comme le chiffre d'affaires ou le nombre de travailleurs pour des entreprises. L'étalement dans l'espace des variables produit alors une sorte de stratification multivariée.

Il existe des techniques simples pour étaler les points dans l'espace. La plus évidente, utilisée dans beaucoup de monitorings environnementaux, consiste sélectionner un échantillon systématique. Une alternative au tirage systématique est la stratification. On découpe l'espace en strates et on sélectionne un très petit nombre d'unités dans chaque strate. Dans l'exemple présenté dans la Figure 4, on sélectionne 64 unités dans une grille de 40 × 40 = 1600 points au moyen d'un plan systématique et d'un plan stratifié avec une seule unité par strate.

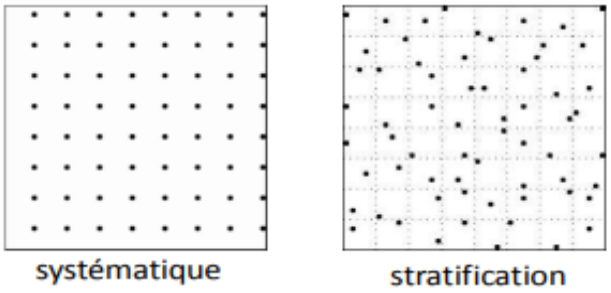


Fig.4. : Dans une grille de 40×40, on sélectionne un échantillon systématique et un échantillon stratifié avec une unité par strate.

Ces deux méthodes ne peuvent cependant pas être appliquées dans toutes les situations. Par exemple, si les unités sont disposées de manière irrégulière dans l'espace, il n'est pas possible de procéder un tirage systématique. Dans ce cas, il est aussi difficile de construire des strates de même taille. De même, le tirage systématique à probabilités inégales (Madow, 1979) ne se généralise pas à deux dimensions. Il est donc nécessaire de considérer d'autres méthodes pour les cas plus généraux.

II.2. Les méthodes

II.2.1. Tessellation aléatoire stratifiée généralisée

La méthode de tessellation (ou pavage) aléatoire stratifiée généralisée (Generalized Random Tessellation Stratified, GRTS) a été proposée par Stevens Jr. & Olsen (2004). Elle permet de sélectionner un échantillon étalé dans un espace avec des probabilités d'inclusion égales ou inégales (voir également Stevens Jr. & Olsen, 1999, 2003; Theobald et al., 2007; McDonald, 2016; Pebesma & Bivand, 2005).

L'objectif de la méthode consiste trier les points selon un ordre (en une dimension) qui respecte les proximités de l'espace en deux dimensions. Pour ce faire, on utilise une fonction quadrant récursive comme le montre la Figure 2. Dans cet exemple, l'espace est divisé en quatre carrés. Chaque carré est nouveau divisé par quatre carrés et ainsi de suite jusqu'à ce qu'il y ait au maximum une unité par carré. Chaque carré a ainsi un label. Par exemple, la cellule grisée de la Figure 5 possède le label (2,3,0).

Stevens Jr. & Olsen (2004) suggèrent ensuite de réaliser une permutation aléatoire l'intérieur de chaque carré pour tous les niveaux. La Figure 6 montre quatre exemples de permutations l'intérieur des carrés.

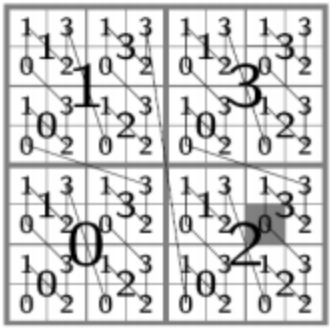


Fig.5. Fonction quadrant récursive utilisée pour la méthode GRTS avec trois subdivisions.

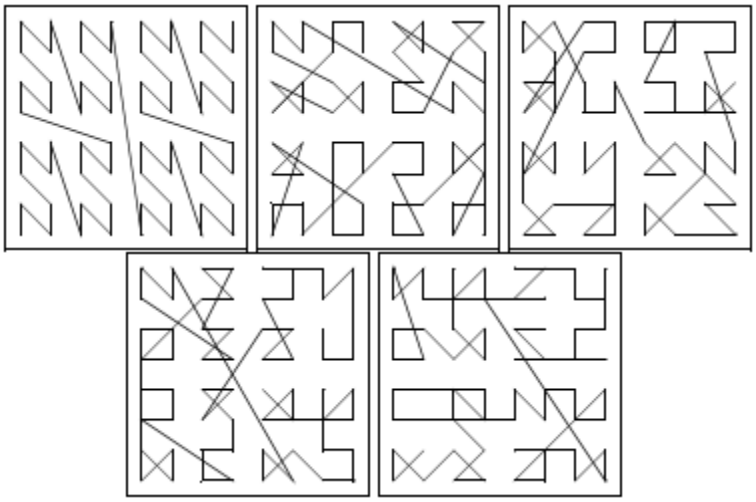


Fig. 6. Fonction d'origine avec quatre permutations

Après avoir sélectionné une ces permutations au hasard, on peut ordonner les carrés et donc les unités. Il y a en effet au maximum une unité par carré. Deux unités qui se suivent sur une de ces courbes sont le plus souvent proches l'une de l'autre dans l'espace. L'échantillon est ensuite sélectionné en utilisant un tirage systématique sur les unités ordonnées. L'échantillon est ainsi relativement bien étalé.

II.2.2. Utilisation de la méthode du voyageur de commerce

Dickson & Tillé (2016) ont proposé plus simplement de calculer le plus court chemin entre les unités de la population. Le calcul de ce chemin est connu comme le problème du voyageur de commerce. Le voyageur de commerce doit passer par un ensemble de villes et revenir la ville de départ tout en minimisant la distance parcourue. Le problème du voyageur de commerce est un problème difficile pour lequel il n'existe pas d'algorithme permettant de trouver avec certitude la solution optimale sans énumérer tous les chemins possibles. En effet, le nombre d'itinéraires possibles est égal $(V - 1)!/2$, où V est le nombre de villes. Il est donc impossible d'énumérer tous les chemins dès que V est grand. Il existe cependant plusieurs algorithmes permettant d'obtenir au moins un minimum local. Quand un itinéraire court est identifié, on peut alors appliquer un tirage systématique à probabilités égales ou inégales.

II.2.3. La méthode du pivot locale

Une approche également très simple a été proposée par Grafström et al. (2012). Ceux-ci ont proposé d'utiliser la méthode du pivot pour réaliser un l'échantillonnage spatial (Deville & Tillé, 1998). À chaque étape, on sélectionne deux unités très proches l'une de l'autre. Ensuite, la méthode du pivot est appliquée sur ces deux unités. Cette méthode est appelée méthode du pivot locale. Si la probabilité d'une de ces deux unités est augmentée, la probabilité de l'autre est diminuée et réciproquement. Ceci induit

une répulsion entre les unités voisines et l'échantillon résultant est ainsi bien étalé.

Grafström et al. (2012) ont proposé plusieurs variantes de cette méthode qui donnent des résultats très similaires. Ce variantes ne diffèrent que dans la manière de sélectionner deux unités voisines dans la population et sont implémentées dans le package R BalancedSampling (Grafström & Lisic, 2016).

II.2.4. La méthode du cube locale

La méthode du cube locale, proposée par Grafström & Tillé (2013), est une extension de la méthode du pivot locale. Cet algorithme permet d'obtenir un échantillon qui est la fois étalé géographiquement et équilibré sur des variables auxiliaires. Cette méthode, appelée méthode du cube locale, consiste lancer la phase de vol de la méthode du cube (Deville & Tillé, 2004), sur un sous-ensemble de J +1 unités voisines, où J est le nombre de variables auxiliaires sur lesquelles on veut équilibrer l'échantillon. Après cette étape, les probabilités d'inclusion sont mises jour de sorte qu'une des J + 1 unités a sa probabilité d'inclusion mise à 0 ou 1, et que les équations d'équilibrage sont toujours satisfaites.

Dans ce groupe de J unités, lorsqu'une unité est sélectionnée, elle diminue les probabilités d'inclusion des J autres unités du groupe. Lorsqu'une unité est définitivement exclue de l'échantillon, elle augmente les probabilités d'inclusion des J autres unités du groupe. Par conséquent, cela induit une corrélation négative dans la sélection des unités voisines, ce qui étale l'échantillon.

La Figure 7 contient un exemple d'échantillonnage de 64 points dans une grille de 40 × 40 = 1600 points au moyen d'un plan simple, de la méthode du pivot locale et de la méthode du cube locale. Les échantillons ont été sélectionnées grâce aux package BalancedSampling et SDraw du langage R (voir Grafström & Lisic, 2016; McDonald, 2016).

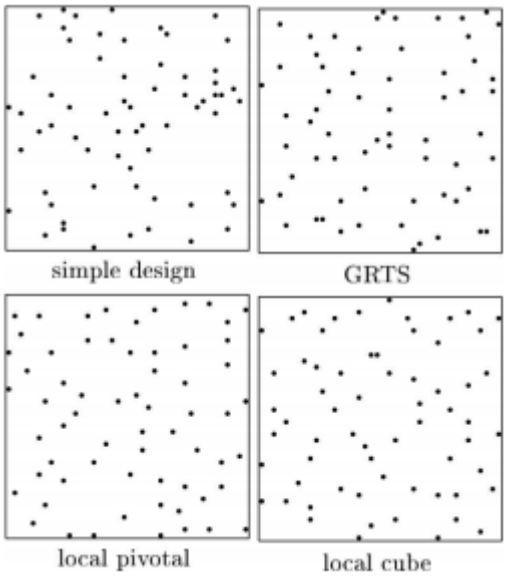


Fig. 7. Échantillonnage de 64 points dans une grille de 40×40 = 1600 points au moyen d'un plan simple, de la méthode du pivot locale et de la méthode du cube locale.

Bibliographie

Tillé, Y. (2018). ÉCHANTILLONNAGE SPATIAL : L'ÉTAT DE L'ART. Université de Neuchâtel, Institut de Statistique. 13es Journees de methodologie statistique de l'Insee (JMS) / 12-14 juin 2018 / PARIS. 7p.