ÉCHANTILLONNAGE SPATIAL : L'ÉTAT DE L'ART

Yves TILLÉ (*)

(*) Université de Neuchâtel, Institut de Statistique

yves.tille@unine.ch

Mots-clés : Autocorrélation, échantillonnage, étalement, équilibrage, probabilités d'inclusion

Résumé

L'échantillonnage spatial est particulièrement important pour la statistique environnementale. Cependant, son champ d'application peut être étendu aux cas où la distance n'est pas géographique. Par exemple, on peut calculer des distances entre des entreprises en se basant sur leurs caractéristiques comme le chiffre d'affaires, le bénéfice ou le nombre de travailleurs. Lorsque deux unités sont proches, elles sont en général similaires, ce qui induit une dépendance spatiale entre les unités. Si l'on sélectionne deux unités proches dans l'échantillon, on aura tendance à récolter une information partiellement redondante. On a donc intérêt à étaler dans l'espace les unités échantillonnées. Nous faisons les points sur les nombreuses méthodes proposées pour étaler des points dans l'espace. Nous proposons aussi une nouvelle méthode d'échantillonnage spatiale.

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 1, 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.

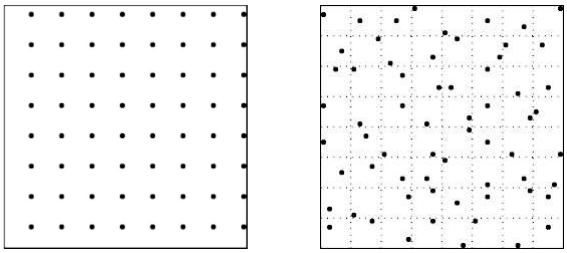

Fig. 1: Dans une grille de 40×40, on sélectionne un échantillon systématique et un échantillon stratifié avec une unité par strate.

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.

2 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 2 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 3 montre quatre exemples de permutations l'intérieur des carrés.

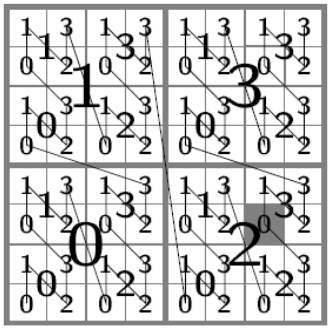

Fig. 2: Fonction quadrant récursive utilisée pour la méthode GRTS avec trois subdivisions.

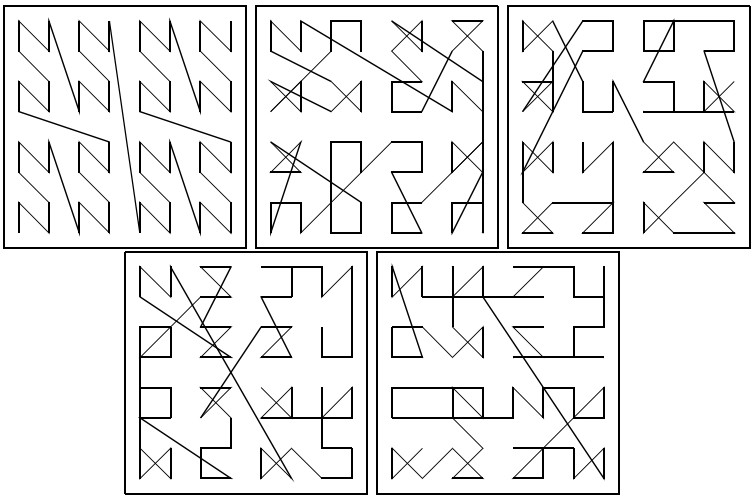

Fig. 3: Fonction d'origine avec quatre permutations

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.

3 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.

4 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).

5 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 4 contient un exemple d'échantillonnage de 64 points dans une grille de $40 \times 40 = 1600$ 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).

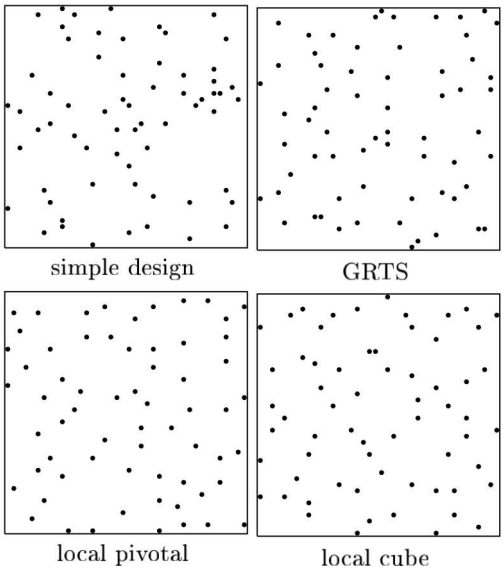

Fig. 4: Échantillonnage de 64 points dans une grille de 40×40 = 1600 points au moyen d'un plan simple, de la méthode du pivot locale et de la méthode du cube locale.

6 Mesures d'étalement

Un polygone de Voronoï centré sur un point *k* est l'ensemble des points qui sont plus proches de *k* que de tout autre point. La Figure 5 montre les polygones de Voronoï autour des points sélectionnés dans les six échantillons présentés dans les Figures 1 et 4. Pour ces exemples, les aires des polygones de Voronoï sont moins dispersées pour les échantillons mieux étalés.

À partir de cette idée, Stevens Jr. & Olsen (2004) ont proposé d'utiliser les polygones de Voronoï pour mesurer l'étalement de l'échantillon. Pour chaque point *k* sélectionné dans l'échantillon, on calcule v_k la somme des probabilités d'inclusion de tous les points de la population qui sont plus proches de *k* que de tous les autres points de l'échantillon. La mesure v_k est donc la somme des probabilités d'inclusion des unités se trouvant dans le polygone de Voronoï centré sur l'unité *k*. Si l'échantillon est de taille fixe *n*,

$$\frac{1}{n}\sum_{k\in U} v_k a_{sk} = \frac{1}{n}\sum_{k\in U} \pi_k = 1 .$$

Pour mesurer l'étalement de l'échantillon $\mathbf{a}_s$, on calcule ensuite simplement la variance des v_k :

$$B(\mathbf{a}_s) = \frac{1}{n}\sum_{k\in U} a_{sk}(v_k - 1)^2 .$$

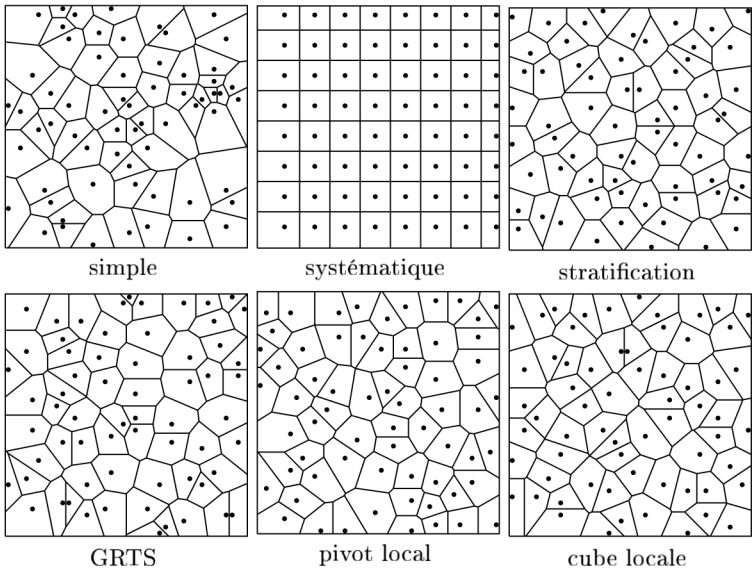

Fig. 5: Échantillonnage de 64 points dans une grille de 40 × 40 = 1600 et polygones de Voronoï. Applications aux plans simples, systématique, stratifié, GRTS, la méthode du pivot locale et la méthode du cube locale.

Plus récemment, Tillé et al. (2018) ont proposé une mesure basée sur l'indice de Moran (1950) d'autocorrélation spatiale. Tillé et al. (2018) ont modifié l'indice de Moran afin qu'il soit strictement compris entre -1 et 1. Il peut donc être interprété comme un coefficient de corrélation. On considère le vecteur d'indicateur de présence dans l'échantillon $\mathbf{a}_s$. Ensuite, on calcule sa corrélation avec le vecteur de ses moyennes locale. La moyenne locale de l'unité *k* est la moyenne des $1/\pi_k - 1$ plus proches valeurs de *k*. Plus précisément, on construit d'abord une matrice de proximité $\mathbf{W} = (w_{k\cdot})$. Pour chaque unité *k* on calcule $q_k = 1/\pi_k - 1$. Puis, on définit

$$w_{k\ell} = \begin{cases} 0 & \text{si } \ell = k \\ 1 & \text{si } \ell \text{ est parmi les } \lfloor q_k \rfloor \text{ plus proches voisins de } k \\ q_k - \lfloor q_k \rfloor & \text{si } \ell \text{ est le } \lfloor q_k \rfloor \text{ème plus proche voisin de } k \\ 0 & \text{sinon.} \end{cases}$$

On calcule ensuite le vecteur contenant *N* fois la moyenne pondérée

$$\bar{\mathbf{a}}_w = \mathbf{U}\frac{\mathbf{a}^\top \mathbf{W}\mathbf{U}}{\mathbf{U}^\top \mathbf{W}\mathbf{U}},$$

où **U** est un vecteur de *N* uns.

Enfin, on définit

$$I_B = \frac{(\mathbf{a}-\bar{\mathbf{a}}_w)^\top \mathbf{W}(\mathbf{a}-\bar{\mathbf{a}}_w)}{\sqrt{(\mathbf{a}-\bar{\mathbf{a}}_w)^\top \mathbf{D}(\mathbf{a}-\bar{\mathbf{a}}_w)\ \ (\mathbf{a}-\bar{\mathbf{a}}_w)^\top \mathbf{B}(\mathbf{a}-\bar{\mathbf{a}}_w)}}.$$

où **D** est une matrice diagonale qui contient les

$$w_{k\cdot} = \sum_{\ell \in U} w_{k\ell}$$

sur sa diagonale,

$$\mathbf{A} = \mathbf{D}^{-1}\mathbf{W} - \frac{\mathbf{U}\mathbf{U}^\top \mathbf{W}}{\mathbf{U}^\top \mathbf{W}\mathbf{U}}$$

et

$$\mathbf{B} = \mathbf{A}\mathbf{D}\mathbf{A}^\top = \mathbf{W}^\top \mathbf{D}^{-1}\mathbf{W} - \frac{\mathbf{W}^\top \mathbf{U}\mathbf{U}^\top \mathbf{W}}{\mathbf{U}^\top \mathbf{W}\mathbf{U}}.$$

L'indice I_B peut s'interpréter comme une corrélation entre l'échantillon et sa moyenne locale. Il vaut -1 si l'échantillon est parfaitement étalé et est proche de 1 si tous les points sont agglomérés. Il est proche de 0 pour un plan simple sans remise.

Le Tableau 1 contient les mesures d'équilibrage spatial $B(\mathbf{a}_s)$ basées sur les polygones de Voronoï et les indices de Moran modifié I_B pour six plans de sondage. Pour la méthode du cube locale, les variables d'équilibrage sont les coordonnées des points et les carrés de ces coordonnées. Le tableau confirme que le plan systématique est le plus étalé. La méthode du pivot locale étale aussi bien que la méthode GRTS.

Tab. 1: Moyennes des mesures d'équilibrage spatial basées sur les polygones de Voronoï $B(\mathbf{a}_s)$ et indices de Moran modifié I_B pour six plans de sondage sur 1000 simulations

Plan	Équilibrage spatial	Indice I_B
Plan simple	0.297	-0.006
Systématique	0.041	-0.649
Stratification	0.063	-0.167
GRTS	0.063	-0.218
Pivot local	0.058	-0.210
Cube local	0.058	-0.210

7 Nouvelle méthode

Aucune des méthodes proposées ne permet d'obtenir un plan composé d'échantillons périodiques quand on sait que ce plan existe. Les échantillons périodiques peuvent être construits sur des grilles quand les inverses des probabilités sont constants et égaux un nombre entier. Nous montrerons plusieurs exemples de tels plans. Ensuite, nous proposerons une nouvelle méthode qui échantillonne en étalant mieux les points que toutes les méthodes existantes et qui permet de construire des plans de sondages périodiques quand ces plans existent.

Bibliographie

[1] Deville, J.-C. & Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* 85, 89-101.

[2] Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* 91, 893-912.

[3] Dickson, M. M. & Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics* 31, 1359-1372.

[4] Grafström, A. & Lisic, J. (2016). *BalancedSampling: Balanced and spatially balanced sampling*. R package version 1.5.2.

[5] Grafström, A. & Lundström, N. L. P. (2013). Why well spread probability samples are balanced? *Open Journal of Statistics* 3, 36-41.

[6] Grafström, A., Lundström, N. L. P. & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics* 68, 514-520.

[7] Grafström, A. & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 14, 120-131.

[8] Madow, W. G. (1949). On the theory of systematic sampling, II. *The Annals of Mathematical Statistics*, 333-354.

[9] McDonald, T. (2016). *SDraw: Spatially Balanced Sample Draws for Spatial Objects*. R package version 2.1.3.

[10] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* 37, 17-23.

[11] Pebesma, E. J. & Bivand, R. S. (2005). *Classes and methods for spatial data in R*. R News 5, 9-13.

[12] Stevens Jr., D. L. & Olsen, A. R. (1999). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics* 4, 415-428.

[13] Stevens Jr., D. L. & Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics* 14, 593-610.

[14] Stevens Jr., D. L. & Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99, 262 -278.

[15] Theobald, D. M., Stevens Jr., D. L., White, D. E., Urquhart, N. S., Olsen, A. R. & Norman, J. B. (2007). Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environmental Management* 40, 134-146.

[16] Tillé, Y., Dickson, M. M., Espa, G. & Giuliani, D. (2018). Measuring the spatial balance of a sample: A new measure based on the Moran's *I* index. *Spatial Statistics* 23, 182-192.