Analyse Factorielle et Sondage - Utilisation de Méthodes d'Échantillonnage Spatial

Ronan Le Gleut [1]

[1] *Insee, Département des Méthodes Statistiques, 75014 Paris, ronan.le-gleut@insee.fr*

Résumé. La propriété d'équilibrage spatial d'un échantillon est importante dans la mesure où elle permet de limiter les problèmes liés à l'auto-corrélation spatiale. Contrairement au cadre usuel de l'échantillonnage spatial, nous allons comparer dans cet article les performances de différentes méthodes d'échantillonnage appliquées à un plan ou un espace factoriel. L'analyse factorielle est effectuée sur des variables présentes dans la base de sondage et corrélées à la variable d'intérêt de l'enquête. Le fait de tirer de façon spatialement équilibrée dans un plan factoriel conduit à réduire assez fortement la variance des estimateurs des variables d'équilibrage et de la variable d'intérêt, sans pour autant accorder un poids trop important à des variables d'équilibrage corrélées entre elles. Cette procédure permet également de mieux restituer la distribution de ces variables en sélectionnant des individus avec des caractéristiques variées. Nous présenterons également une nouvelle méthode de tirage d'échantillons spatialement répartis implémentée sous **R**. Cette procédure se base sur l'algorithme de Tessellation de la méthode GRTS (en améliorant le temps d'exécution) et sur la méthode du Pivot. Les performances de cette méthode sont comparées à d'autres méthodes d'échantillonnage spatial en termes d'EQM et de degré d'équilibrage spatial.

Mots-clés. Analyse factorielle, échantillonnage spatial, méthode du Pivot.

Abstract. The degree of spatial balance of a sampling design is very important in order to limit a lack of efficiency due to spatial auto-correlation between units. The aim of this article is to compare the performance of different sampling methods in a factor plane or a factor space rather then in a geographical space as usual. The factor analysis is done on variables available in the sampling frame which are correlated with the target variable of the survey. In this context, the variance of the estimators for the balancing variables and for the variable of interest is reduced, without according a major importance to correlated balancing variables. With this procedure, the estimated distribution of these variables is also closer to the real one *via* the selection of units with varying characteristics. We will also present a new method of spatial sampling implemented in **R**. This procedure is based on the algorithm of the GRTS method in order to tessellate the space (improving computational time), and on the Pivotal method to select a sample. The performance of this method is compared to other methods of spatial sampling in terms of MSE and degree of spatial balance.

Keywords. Factor analysis, Pivotal method, spatial sampling.

1 Introduction

L'intérêt de l'échantillonnage spatial a été démontré dans de nombreux articles scientifiques, que ce soit pour des études environnementales, de géologie ou de géographie, mais également dans le cadre de la statistique publique avec des enquêtes auprès des entreprises ou des ménages, voir par exemple Favre-Martinoz et Merly-Alpa (2016).

La propriété d'équilibrage spatial est importante dans la mesure où elle permet de sélectionner des unités éloignées les unes des autres. En effet, les unités sont souvent auto-corrélées spatialement, ce qui signifie que deux unités proches ont tendance à être similaires au sens de certaines variables (et inversement pour des unités éloignées). Tirer des unités proches aurait ainsi des conséquences néfastes en termes de précision pour les variables auto-corrélées spatialement.

Cet article a pour objectif de comparer différentes méthodes d'échantillonnage spatial dans un cadre non spatial, mais pour lequel les problèmes d'auto-corrélation existent. En effet, au lieu d'appliquer les différentes méthodes dans un espace géographique avec des coordonnées (x, y), nous exploiterons les résultats d'une analyse factorielle effectuée sur des variables corrélées à notre variable d'intérêt afin de tirer un échantillon « spatialement » réparti sur un plan factoriel. Cette procédure a ainsi pour objectif de s'approcher de la notion d'équilibrage sur des variables auxiliaires introduite par Deville et Tillé (2004). Nous évaluerons ainsi les performances des différentes méthodes en termes de variance sur la variable d'intérêt ainsi que sur les variables d'équilibrage.

En complément du domaine d'application non spatial, nous présenterons une nouvelle procédure d'échantillonnage spatial à l'aide de la méthode du Pivot. Cette procédure reprend les idées de la méthode *Generalized Random-Tessellation Stratified* (GRTS) proposée par Stevens et Olsen (2004) tout en diminuant le temps d'exécution nécessaire au tirage d'un échantillon. Les performances de cette méthode en termes d'équilibrage spatial et de variance seront comparées avec d'autres méthodes d'échantillonnage spatial.

2 Analyse factorielle et Sondage

Le tirage équilibré, est une procédure dont le but est de fournir un échantillon respectant les deux contraintes suivantes, au moins approximativement :

— les probabilités d'inclusion dans l'échantillon sont respectées ;
— l'échantillon est équilibré sur p variables auxiliaires disponibles dans la base de sondage. Autrement dit, les estimateurs d'Horvitz-Thompson des totaux des variables d'équilibrage sont égaux aux totaux de ces variables dans la population.

L'intérêt de tirer un échantillon équilibré réside dans le fait que si ces variables auxiliaires sont corrélées à la variable d'intérêt y, alors la variance d'échantillonnage de l'estimateur d'Horvitz-Thompson du total de y sera réduite. De plus, la variance des variables d'équilibrage sera faible, voir nulle si l'équilibrage est exact.

L'objectif est ici de s'approcher de l'équilibrage total proposé par la méthode du cube, voir Deville et Tillé (2004), en sélectionnant un échantillon « spatialement » réparti dans un plan factoriel ou un espace factoriel si les deux premiers axes ne résument pas suffisamment l'information. De la même façon que pour la méthode du cube, les variables utilisées pour l'analyse factorielle (ACP pour des variables quantitatives, AFM pour des données mixtes, etc.) sont corrélées à la variable d'intérêt.

Le fait de tirer « équilibré » dans un plan factoriel permet ainsi de sélectionner des individus avec des caractéristiques variées, sans pour autant apporter un poids trop important à des variables d'équilibrage fortement corrélées entre elles (celles-ci étant souvent résumées sur un seul axe). Ainsi, la variance des estimateurs des variables d'équilibrage et de la variable d'intérêt sera fortement réduite, d'autant plus pour l'estimation d'un quantile (la distribution des variables étant mieux restituée).

3 La méthode du Pivot par Tessellation

La méthode du Pivot est une méthode d'échantillonnage permettant de sélectionner un échantillon avec des probabilités d'inclusion inégales, voir Deville et Tillé (1998). A chaque étape de la méthode, les probabilités sont mises à jour pour deux unités en lice et l'une au moins de ces deux unités est sélectionnée ou rejetée. La méthode est simple à mettre en oeuvre, et des résultats théoriques tels que le théorème central limite s'appliquent, voir Chauvet et Le Gleut (2017).

L'algorithme de Tessellation présenté dans cette partie s'inspire fortement de Stevens et Olsen (2004). Afin d'obtenir un découpage de l'espace efficace, nous nous appuyons sur la décomposition binaire d'un nombre qui est faite par défaut dans le logiciel **R**. L'idée est la suivante :

(1) Les coordonnées x_i et y_i d'une unité i sont projetées sur un carré $[0, 2^{31} - 1] \times [0, 2^{31} - 1]$, ce qui permet d'exprimer chaque coordonnée en base 2 à l'aide d'un code Bit sur 31 positions.

(2) Le carré est ensuite divisé en quatre « descendants » de même taille grâce à la première position des codes Bit $x_i[1]$ et $y_i[1]$ valant 0 ou 1 :
   — si $(x_i[1]y_i[1])_2 = (00)_2 = 0 \rightarrow$ l'unité i appartient au carré en bas à gauche d'adresse 0,
   — si $(x_i[1]y_i[1])_2 = (01)_2 = 1 \rightarrow$ l'unité i appartient au carré en haut à gauche d'adresse 1, etc ...

La transformation de l'étape (1) permet d'effectuer la décomposition du carré en « descendants » de l'étape (2) directement à 31 niveaux, ce qui conduit à obtenir un découpage en $4^{31} \approx 4.6\ 10^{18}$ petits carrés. L'adresse d'une unité est la concaténation de toutes les adresses successives des carrés auxquels elle a appartenu au cours de la décomposition. La base de sondage est alors triée selon les « adresses » de chaque unité afin d'obtenir un

« chemin » parcourant l'espace à deux dimensions. Un échantillon spatialement équilibré est obtenu en utilisant la méthode du Pivot sur le fichier trié.

La Figure 1 présente un exemple de découpage et d'attribution des adresses, en se limitant aux trois premiers niveaux de décomposition (code Bit sur trois positions). Les valeurs maximale et minimale des coordonnées projetées sont donc 0 (000) et 7 (111). Dans cet exemple, les coordonnées projetées de l'unité i sont $(x_i, y_i) = (2, 6)$, ce qui conduit à une décomposition binaire de $(010, 110)$. La première position du code Bit indique que l'unité i appartient au carré à gauche ($x_i[1] = 0$) et en haut ($y_i[1] = 1$), i.e. au descendant n°1 ($(01)_2$). La deuxième position du code Bit indique que, dans ce premier descendant, l'unité i apparitient au carré en haut ($y_i[2] = 1$) à droite ($x_i[2] = 1$), i.e. au descendant n°3 ($(11)_2$) du descendant n°1, et ainsi de suite ...

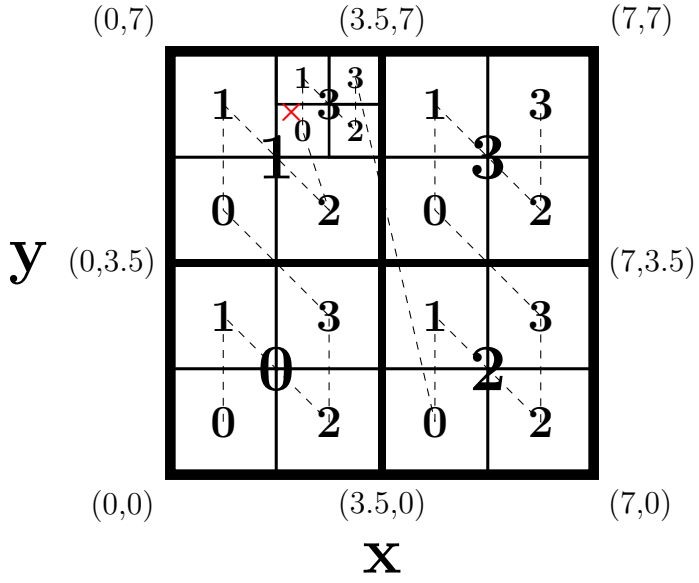
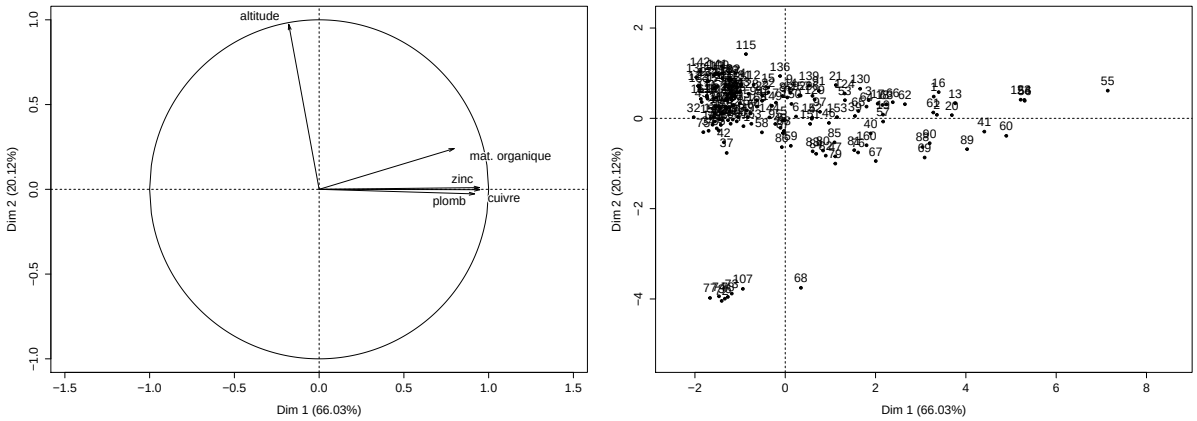


FIGURE 1 – Trois premières divisions du carré. L'adresse associée à l'unité i (représentée par une croix rouge) dont les coordonnées sont $(x_i, y_i) = (2, 6) = (010, 110)$ est « 130 ».

4 Comparaison de méthodes

Afin d'évaluer les performances de la méthode du Pivot par Tessellation, nous utilisons le jeu de données « Meuse » disponible dans le *package* `gstat` du logiciel **R**. Ce jeu de données contient les concentrations en métaux lourds de N=164 localisations proches du village Stein (Pays-Bas).

La variable d'intérêt est la concentration en cadmium, et les variables d'équilibrage utilisées dans l'analyse factorielle (corrélées à la variable d'intérêt, voir Section 2) sont les



(a) (b)

FIGURE 2 – Projection des variables (2a) et des individus (2b) sur le premier plan factoriel représentant plus de 85% de l'inertie totale.

concentrations en cuivre, plomb, zinc, ainsi que le pourcentage de matière organique et l'altitude. La Figure 2 représente la projection des variables (2a) et des individus (2b) sur le premier plan factoriel issu de l'ACP réalisée sur les variables d'équilibrage. C'est dans cet espace que les méthodes d'échantillonnage spatial vont être comparées.

Nous allons donc comparer la méthode du Pivot par Tessellation avec d'autres méthodes telles que le Pivot local introduit par Grafström, Lundström et Schelin (2012), la méthode GRTS de Stevens et Olsen (2004), les méthodes d'échantillonnage ordonnées à l'aide du problème du voyageur de commerce détaillées dans l'article de Dickson et Tillé (2016), ou encore la méthode du cube spatialement équilibré proposée par Graftröm et Tillé (2013).

Les simulations ont consisté à répliquer 10 000 tirages d'échantillons de taille n=50 pour chacune des méthodes citées précédemment, avec des probabilités d'inclusion égales ou proportionnelles à la concentration en cuivre. Pour toutes ces méthodes, nous cherchons à évaluer deux aspects :

(1) l'erreur quadratique moyenne (EQM) de l'estimateur d'Horvitz-Thompson pour l'estimation d'un total ou d'un quantile de la variable d'intérêt et des variables d'équilibrage ;

(2) le degré d'équilibrage « spatial » (i.e. sur le premier plan factoriel) de l'échantillon *via* l'approche des polygones de Voronoi suggérée par Stevens et Olsen (2004) ;

5 Résultats

Nous montrerons qu'en termes d'EQM, le fait de tirer « spatialement » équilibré dans un plan factoriel permet de réduire assez fortement la variance pour les variables utilisées

dans l'analyse factorielle ainsi que pour la variable d'intérêt. Pour l'estimation d'une distribution, les méthodes présentent des résultats assez proches.

Concernant le degré d'équilibrage spatial, les méthodes présentent également des résultats à peu près similaires, avec tout de même un meilleur équilibrage pour le Pivot local mettant en lice deux unités les plus proches l'une de l'autre.

6 Discussion

La méthode du Pivot par Tessellation présente plus de flexibilité que la méthode GRTS. D'une part, l'algorithme de Tessellation est dissocié de la partie tirage, ce qui permet par exemple de choisir une autre méthode d'échantillonnage (e.g. systématique, comme GRTS). D'autre part, il est possible de ne pas se limiter à deux dimensions, mais d'effectuer ce découpage sur p variables. Également, tout comme pour la méthode GRTS, il est possible de « randomiser » la Tessellation (i.e. de changer l'ordre d'attribution des adresses à chaque étape du découpage). Enfin, l'un des principaux avantages est le temps d'exécution de la méthode. Par exemple, le fait de sélectionner un échantillon de taille n=100 000 dans une population de taille N = 1 000 000 prend 40 secondes avec cette méthode (codée en **R**), là où les autres méthodes prennent plusieurs heures (GRTS, Pivot local en **C++**, etc.).

Bibliographie

[1] Chauvet, G., et Le Gleut, R. (2017), Asymptotic Results for Pivotal Sampling with Application to Spatial Sampling, *Travail en cours*.

[2] Deville, J. C., et Tillé, Y. (1998), Unequal Probability Sampling Without Replacement Through a Splitting Method, *Biometrika*, 89–101.

[3] Deville, J. C., et Tillé, Y. (2004), Efficient Balanced Sampling : the Cube Method, *Biometrika*, 893-912.

[4] Dickson, M. M., et Tillé, Y. (2016), Ordered spatial sampling by means of the traveling salesman problem, *Computational Statistics, 31*(4), pp. 1359–1372.

[5] Favre-Martinoz, C. et Merly-Alpa, T. (2016), Utilisation des Méthodes d'Échantillonnage Spatialement Équilibré pour le Tirage des Unités Primaires des Enquêtes Ménages de l'Insee, *9ème Colloque Francophone sur les Sondages*, Gatineau.

[6] Grafström, A., Lundström, N. L. et Schelin, L. (2012), Spatially Balanced Sampling through the Pivotal Method, *Biometrics, 68*(2), 514–520.

[7] Grafström, A., et Tillé, Y. (2013), Doubly Balanced Spatial Sampling with Spreading and Restitution of Auxiliary Totals, *Environmetrics, 24*(2), 120-131.

[8] Stevens Jr, D. L. et Olsen, A. R. (2004), Spatially Balanced Sampling of Natural Resources, *Journal of the American Statistical Association, 99*(465), 262–278.