Peut-on mobiliser la dimension spatiale à l'étape de redressement d'une enquête ?

Laurie Leterrier [1] & Thomas Merly-Alpa [2]

[1] *ENSAI, France, laurie.leterrier@eleve.ensai.fr*
[2] *Ined, France, thomas.merly-alpa@ined.fr*

Résumé. La prise en compte de la dimension spatiale (proximité, éloignement) dans le tirage des échantillons a pour conséquence une amélioration de la précision des estimations pour des variables ayant une autocorrélation spatiale. Cette communication vise à explorer plusieurs méthodes (calage, imputation par les plus proches voisins, lissage spatial) de mobilisation ex-post de la position relative des individus échantillonnés. Une étude par simulations est réalisée pour estimer le gain en variance attendu avec l'application de ces méthodes. Une discussion sur les limites (biais, réplicabilité) est réalisée.

Mots-clés. Redressements, calage, imputation, statistiques spatiales ...

Abstract. Taking account of the spatial dimension (proximity, distance) in the drawing of samples results in an improvement in the accuracy of estimates for variables with spatial autocorrelation. The aim of this paper is to study several methods (calibration, nearest neighbour imputation, spatial smoothing) using spatial dispersion ex-post. A simulation study is carried out to estimate the expected gain in variance with the application of these methods. Limitations (bias, replicability) are discussed.

Keywords. Reweighting, calibration, imputation, spatial statistics ...

1 Introduction

La question du territoire et de la définition des zones de collecte a toujours été centrale dans les enquêtes par entretien. Cependant, depuis quelques années, et notamment avec le développement du sondage dit "doublement équilibré" (Grafström et Tillé, 2013), la prise en compte de la dimension spatiale dans l'élaboration d'un plan de sondage pour une enquête en sciences sociales (comme l'échantillon-maître français, Chevalier et al., 2022) est en plein renouveau. Ces méthodes reposent sur la première loi de la géographie de Tobler ("Tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés", Tobler, 1970), et visent ainsi à minimiser les situations dans lesquelles deux individus géographiquement proches sont conjointement sélectionnés dans un échantillon. [1]

[1] Ce travail a été réalisé à l'Ined durant l'été 2023 par Laurie Leterrier dans le cadre de son stage de deuxième année de l'ENSAI.

Lorsque la dimension spatiale est utilisée lors de l'échantillonnage, les individus sélectionnés sont géographiquement dispersés. Ces échantillons réduisent alors considérablement la variance des estimations, engendrant ainsi un gain important de précision dans l'estimation.

La dimension spatiale pourrait également être appréhendée au moment de l'estimation, lorsque des poids de sondage propres à chaque individu de l'échantillon sont recalculés après l'étape d'échantillonnage. L'objectif de cette communication est d'évaluer s'il est possible d'utiliser cette dimension au moment de l'étape de redressement de l'enquête. Nous proposons ici d'explorer plusieurs méthodes : le calage, l'imputation par le plus proche voisin et le lissage spatial.

2 Méthodes proposées

2.1 Calage sur marges

Il est possible de prendre en compte la dimension spatiale après l'échantillonnage en effectuant un calage sur marges (Deville et Särndal, 1992) avec des variables liées à la géographie. Cette idée n'est pas nouvelle : dans la plupart des redressements d'enquêtes une variable de niveau géographique est utilisée (région, département, zone d'activité, etc.). Il serait possible d'étendre ce cadre au-delà des zonages administratifs en utilisant la latitude et la longitude comme variables de calage. Il est également possible d'effectuer un calage sur des polynômes de la latitude et la longitude. Cela permet de prendre en compte des aspects non linéaires de la répartition géographique.

Cependant, la littérature scientifique sur la statistique spatiale recommande plutôt d'utiliser la notion de voisinage. Nous proposons deux méthodes pour cela.

2.2 Imputation par le plus proche voisin

Une autre manière de prendre en compte la dimension spatiale est d'attribuer, pour un individu de la population non sélectionné dans l'échantillon, la valeur de la variable d'intérêt Y de l'individu de l'échantillon dont il est le plus proche géographiquement. Cette méthode revient à attribuer à chaque individu de l'échantillon un poids égal au nombre d'individus dans la population qui sont plus proches de cet individu que des autres individus échantillonnés.

2.3 Lissage spatial

Enfin, une dernière méthode envisagée est le lissage spatial, en suivant l'exemple de Cowling et al., 1997. Le lissage spatial consiste à calculer la moyenne de la variable d'intérêt Y au voisinage d'un individu non échantillonné, que l'on notera individu i, affectée de coefficients dépendant de la distance. Plus les individus de l'échantillon seront proches de l'individu i, plus les coefficients qui leur seront attribués seront importants. La variable d'intérêt

Y imputée par lissage spatial pour l'individu i sera ainsi calculée grâce à l'estimateur de Nadaraya-Watson :

$$\widehat{Y_i} = \frac{\sum_{j=1}^{n} K(\frac{x_i - x_j}{h}) Y_j}{\sum_{l=1}^{n} K(\frac{x_i - x_l}{h})}.$$

où x représente la position géographique des individus, Y la variable d'intérêt.

Ces coefficients sont définis à l'aide d'une fonction appelée *noyau*. Dans ce travail, les lissages suivants sont effectués à l'aide du noyau quartique, défini par :

$$K(u) = \frac{15}{16}(1 - u^2)^2 \mathbb{1}_{|u| \leq 1}.$$

D'autres noyaux existent afin d'appréhender le voisinage, mais, selon Scott, 1992, le choix du noyau importe peu. D'autres simulations ont été effectuées avec d'autres noyaux et confirment ce point.

Pour définir le voisinage d'un individu, on utilise un paramètre de lissage, noté h appelé bande passante. Ce dernier va contrôler la taille du voisinage. Plus il sera grand, plus le nombre d'individus de l'échantillon considérés comme voisins de l'individu i sera grand. Le lissage sera alors plus fort et la variance diminuera. Au contraire, si le paramètre de lissage est faible, peu d'individus seront voisins de l'individu i : le lissage sera plus faible et la variance augmentera. Le choix de la valeur de bande passante retenue correspond alors à un compromis biais-variance : on choisira dans la suite des travaux le paramètre permettant de minimiser l'erreur quadratique moyenne. Ce choix est possible ici car nous sommes dans le cadre de simulations : dans l'application à partir d'un unique échantillon, il ne serait pas possible de déterminer la meilleure valeur de bande passante de cette façon - des méthodes de validation croisée seraient alors à envisager.

Si l'individu i pour lequel on souhaite estimer Y ne possède pas de voisin dans l'échantillon, relativement au paramètre de lissage choisi, on estime Y_i par la moyenne des valeurs de la variable d'intérêt au sein de l'échantillon.

Le lissage spatial peut également être lu comme une étape de repondération. Il s'agit dans ce cas de faire "hériter" les poids des individus non échantillonnés à ceux échantillonnés, en utilisant les coefficients dans la formule de lissage spatial (les individus proches suivant le noyau utilisé hériteront d'une plus grande part du poids).

3 Simulations et résultats

3.1 Simulation des données

On génère une population de taille $N = 1\,000$ pour laquelle on va générer trois caractéristiques. On génère tout d'abord une variable explicative X en suivant une loi normale centrée réduite. On associe ensuite à chaque individu une position dans l'espace : sa latitude et longitude sont générées selon une loi uniforme sur $[0\,;1]$.

Enfin, la génération de la variable d'intérêt Y est plus complexe. Afin d'avoir une variable qui présente de l'autocorrélation spatiale, nous utilisons un modèle spatial autorégressif (SAR), qui peut être généré en suivant la méthode décrite dans Loonis et de Bellefon, 2018, par la formule suivante, où ϵ est un bruit gaussien centré réduit :

$$Y = (1 - \rho W)^{-1} X\beta + (1 - \rho W)^{-1}\epsilon, \ \beta \text{ étant fixé à } 1.$$

ρ contrôle la corrélation spatiale : plus ρ est proche de 1 en valeur absolue, plus cette corrélation est élevée (positivement ou négativement). W est la matrice de voisinage, calculée avec la distance euclidienne à partir des latitudes et longitudes des individus avec un plafond égal à la distance minimale pour laquelle chaque individu possède au moins un voisin. Enfin, afin de garantir des Y positifs, on ajoute $|min(Y)|$ à chaque valeur de Y.

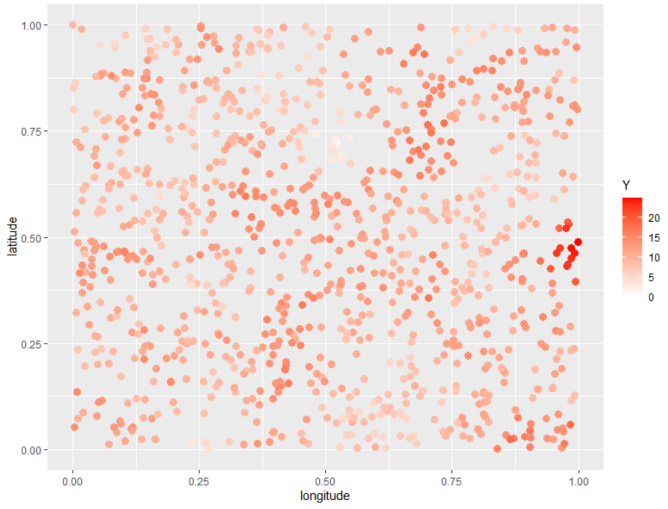

Figure 1: Répartition de la variable Y

Dans notre population, on fixe $\rho = 0.9$. L'indice de Moran permet de calculer l'intensité de l'autocorrélation spatiale. Il est défini pour une variable d'intérêt Y par :

$$I_W = \frac{N}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}, \; i \neq j,$$

avec $W = (w_{ij})_{i,j=1,\ldots,N}$ la matrice de voisinage, telle que $w_{ij} = 1$ si les individus i et j sont considérés comme voisins, $w_{ij} = 0$ sinon, et N la taille de la population.

Dans nos simulations, nous obtenons un indice de Moran de 0.79, ce qui signifie que la corrélation spatiale des données est extrêmement élevée. Ce scénario ne correspond pas à ce qui est usuellement rencontré dans la pratique, mais va permettre d'étudier si les méthodes de redressement proposées peuvent fonctionner ou non dans le cas le plus favorable.

La répartition de la variable Y est représentée sur la figure 1. Les individus proches ont tendance à avoir des valeurs similaires pour Y, ce qui correspond bien à une corrélation spatiale élevée et positive.

3.2 Méthodes d'échantillonnage

Plusieurs échantillons sont tirés au sein de cette population. On tire $nsamp = 10\ 000$ échantillons de taille $n = 50$, soit un taux de sondage de 5%.

Ces échantillons sont tirés selon trois familles de plans de sondages, en utilisant des probabilités de tirages égales :

1. le sondage aléatoire simple sans remise (SRS) ;
2. le cube spatialement équilibré, développé par Grafström et Tillé, 2012, en intégrant la contrainte de taille fixe comme variable d'équilibrage ;
3. le tirage systématique le long d'un chemin parcourant la population. Ce chemin est obtenu en cherchant une solution du problème au voyageur de commerce, tel que décrit par Applegate et al., 2000.

Ces trois plans de sondage ont été sélectionnés afin d'avoir d'une part un cadre de référence, et d'autre part deux dernières techniques permettent d'obtenir des échantillons spatialement dispersés. D'autres plans de sondages mobilisant l'information auxiliaire sur X peuvent être envisagés : par exemple, des variantes avec des probabilités de sélection proportionnelles à X ont été testées et ne modifient pas substantiellement les résultats présentés ici.

3.3 Résultats

Nous allons évaluer la performance des méthodes de redressement proposées (calage, imputation, lissage spatial) sur l'évaluation de deux paramètres :

1. Le total de la variable Y sur la population entière ;
2. Le total de Y au sein de quatre sous-populations définies selon des critères géographiques (découpage en 4 zones du plan).

La performance des méthodes sera évaluée en calculant la variance (et notamment le design effect, c'est-à-dire le ratio entre la variance obtenue et celle obtenue par sondage aléatoire simple sans redressements) et le biais des estimateurs obtenus.

3.3.1 Population totale

Le tableau 1 récapitule l'ensemble des *design effects* pour chaque méthode d'échantillonnage, et pour chaque calcul des poids de sondage.

	SRS	Cube spatialement équilibré	Voyageur de commerce
Sans redressement	1	0.54	0.49
Imputation	0.88	0.48	0.44
Lissage spatial	0.77	0.46	0.42

Table 1: *Design effect* pour chaque méthode étudiée

Les résultats correspondants à la méthode de calage sur latitude, longitudes ou des polynômes de ces variables ne sont pas inclus dans le tableau. Ils conduisent à des précisions sensiblement identiques à la situation sans redressement, et ces méthodes semblent ainsi peu efficaces.

En revanche, les méthodes de redressement basées sur le voisinage conduisent à des variances plus faibles que l'utilisation directe des poids issus de l'échantillonnage. Ce gain est plus important pour un échantillon qui n'est pas initialement spatialement dispersé, comme le sondage aléatoire simple.

Une autre conclusion est que prendre en compte la dimension spatiale lors de l'étape d'échantillonnage est largement plus efficace que de l'utiliser au moment des redressements. C'est un résultat similaire à celui, classique, qui dit que la stratification est plus efficace que la post-stratification.

	SRS	Cube spatialement équilibré	Voyageur de commerce
Sans redressement	0.01%	-0.02%	-0.01%
Imputation	0.21%	0.22%	0.24%
Lissage spatial	0.22%	0.32%	0.34%

Table 2: Biais relatif pour chaque méthode étudiée

Tous les estimateurs ont des biais nuls ou faibles (voir tableau 2). Le lissage spatial, comme indiqué en partie 2.3, conduit à introduire un biais du fait du choix d'une bande passante importante. Cependant, le choix d'un paramètre minimisant l'erreur quadratique moyenne conduit à des biais relativement faibles.

3.3.2 Sous-espaces géographiques

Un autre enjeu peut être celui de l'estimation sur des sous-espaces géographiques, comme discuté par Cowling et al., 1996. Cela permet de réaliser des diffusions de résultats locaux, y compris si la géographie adaptée n'a pas été prise en compte dès l'échantillonnage. Dans nos simulations, nous découpons le plan en quatre régions : Sud-Ouest, Nord-Ouest, Sud-Est et Nord-Est. On calcule ensuite l'estimation du total de la variable d'intérêt sur chacune des quatre régions.

L'imputation par le plus proche voisin et le lissage spatial réduisent considérablement l'écart-type pour chaque région par rapport aux estimations faites sans redressements, le gain étant plus important pour le lissage spatial. Le tableau 3 résume les écarts-types obtenus pour l'une des régions (Sud-Ouest) ; les résultats sont du même ordre sur les quatre régions.

	SRS	Cube spatialement équilibré	Voyageur de commerce
Sans redressement	676	282	260
Imputation	164	131	129
Lissage spatial	100	81	79

Table 3: Ecart-type du total de Y dans la région Sud-Ouest.

Cependant, les deux méthodes d'imputation et de lissage spatial entraînent un biais sur l'estimation du total au sein de chaque sous-région. Ce biais est particulièrement important dans le cadre du lissage spatial (voir Tableau 4).

	SRS	Cube spatialement équilibré	Voyageur de commerce
Sans redressement	0.46%	-0.26%	-0.08%
Imputation	-0.39%	0.35%	0.44%
Lissage spatial	-3.19%	-3.08%	-3.04%

Table 4: Biais relatif dans l'estimation du total de Y dans la région Sud-Ouest.

Ce biais s'explique probablement par deux raisons. D'une part, comme pour l'estimation du total sur la population, le lissage spatial conduit à l'introduction d'un biais via le choix d'une bande passante importante. D'autre part, le biais peut s'expliquer ici par la possibilité d'attribuer par imputation ou lissage des valeurs issues d'une autre région à des individus appartenant à une région prédéfinie. Pour l'illustrer, dans l'hypothèse où des individus avec un Y plus important vivent de l'autre côté de la frontière, les imputations et le lissage conduisent à augmenter mécaniquement l'estimation du total de Y dans la sous-région considérée.

Conclusion

Lorsque la variable d'intérêt est positivement et spatialement autocorrélée, prendre en compte la dimension spatiale dans des enquêtes permet d'augmenter la précision des estimations. Le gain de précision le plus important est obtenu au moment de l'échantillonnage, en sélectionnant un échantillon spatialement dispersé. Si on n'utilise la dimension spatiale que lors de l'estimation, le gain de précision est moins important mais toujours présent.

Ce gain de précision est particulièrement important pour des estimations sur des sous-domaines géographiques, mais il peut conduire à des biais importants. Une étude du compromis biais-variance est donc à réaliser avant de mettre en place ces méthodes.

Les résultats présentés dans ce document restent très exploratoires : les seuls résultats sont sur des simulations où la population est petite, le taux de sondage grand et l'autocorrélation spatiale particulièrement forte. Par ailleurs, la méthode de lissage spatial repose sur des paramètres (noyau, bande passante) qui sont à calibrer finement : la détermination des bons paramètres à partir d'un échantillon reste à étudier. Des travaux complémentaires sur données réelles sont ainsi nécessaires pour conclure sur l'applicabilité de cette méthode.

Bibliographie

Applegate, David, Cook, William et Rohe, André (2000). "Chained Lin-Kernighan for large traveling salesman problems". In : INFORMS Journal on Computing 15(1), p. 82-92.

Chevalier, Martin et al. (2022). "Le renouvellement de l'échantillon-maître des enquêtes auprès des ménages et de l'échantillon de l'enquête Emploi de l'Insee". Insee Méthodes n° 141.

Cowling, Ann et al. (1996). "Application du lissage spatial aux données d'enquête". In : Techniques d'enquête, Statistique Canada 22.2, p. 177-186.

Deville, J.-C. and Särndal, C.-E (1992), "Calibration estimation in survey sampling", Journal of the American Statistical Association, 87, n°418, pp. 375-382.

Grafström, Anton et Yves Tillé (2013). "Doubly balanced spatial sampling with spreading and restitution of auxiliary totals". In : Environmetrics 24(2), p. 120-131.

Loonis, Vincent et Marie-Pierre Bellefon (2018). Manuel d'analyse spatiale. Théorie et mise en oeuvre avec R. Insee Méthodes n°131.

Scott, David W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Series in Probability et Statistics

Tobler, Waldo R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." Economic Geography (Supplement: Proceedings, International Geographical Union. Commission on Quantitative Methods), 46: 234–240. DOI:10.2307/143141.