10. Échantillonnage spatial

CYRIL FAVRE-MARTINOZ, MAËLLE FONTAINE, RONAN LE GLEUT, VINCENT LOONIS
Insee

Résumé

Dans ce chapitre, nous nous intéressons à l'utilisation de l'information géographique dans le contexte de l'échantillonnage. Cette information peut être utilisée à différents moments dans le processus de conception d'un plan de sondage. Dans une grande partie des enquêtes en face-à-face, des plans de sondages à plusieurs degrés sont utilisés afin de réduire les coûts de collecte en concentrant géographiquement les interviews. Un géoréférencement fin des unités statistiques à échantillonner s'avère déterminant pour la constitution des entités à sélectionner au(x) premier(s) degré(s). Cette information géographique peut également être mobilisée au moment de la sélection de l'échantillon, afin d'améliorer l'efficacité statistique de celui-ci en présence d'autocorrélation spatiale positive des variables d'intérêt.

R La lecture préalable des chapitres 2 : "Codifier la structure de voisinage" et 3 : "Indices d'autocorrélation spatiale" est recommandée.

Introduction

Le projet Geostat 2 d'Eurostat (2015-2017) visait à fournir un cadre de référence permettant une production efficiente et une utilisation aisée d'information statistique finement localisée. Concernant les enquêtes par sondage, le rapport final du projet : *A Point-based Foundation for Statistics*, identifie au moins trois phases de la conception d'une enquête qui pourraient bénéficier d'une base de sondage géocodée. Premièrement, en amont, lorsque le mode de collecte est le face-à-face, la connaissance précise de la localisation de l'ensemble des unités statistiques permet la création d'unités primaires [1] (UP). La connaissance des caractéristiques de ces UP facilite la gestion du réseau d'enquêteurs tout en maintenant les qualités statistiques de l'échantillonnage. Deuxièmement, quel que soit le mode de collecte, l'information géographique permet, sous certaines conditions, d'améliorer la précision des estimations en mobilisant des méthodes d'échantillonnage spatial. Troisièmement, lors de la phase de collecte, la connaissance de la localisation des unités statistiques échantillonnées permet de faciliter leur repérage quand la qualité de l'adressage n'est pas suffisante.

Dans ce chapitre nous nous intéressons exclusivement aux deux premiers points. Nous rappelons brièvement en première partie le cadre de la théorie des sondages. Ensuite, la deuxième partie présente une méthode de constitution d'unités primaires ayant un nombre constant d'unités statistiques, tout en ayant une faible étendue géographique. La présentation des différentes méthodes d'échantillonnage spatial constitue la troisième partie, alors que la dernière partie compare leurs propriétés de façon empirique par simulation.

Parmi la riche littérature existant sur le sujet, nous nous appuyons sur, ou orientons le lecteur vers BENEDETTI et al. 2015.

10.1 Généralités

L'objectif de la théorie des sondages est d'estimer la valeur d'un paramètre θ mesuré sur une population U de taille N [2]. Ce paramètre est fonction des valeurs prises par une, ou plusieurs, variable(s) d'intérêt associée(s) à chacun des individus de la population. On note y_i la valeur de la variable d'intérêt y pour l'individu i de U. Le sondeur n'a accès aux y_i que pour une sous-partie de la population appelée échantillon et notée s. Il agrège les valeurs observées sur l'échantillon par une fonction, appelée estimateur, prenant la valeur $\hat{\theta}(s)$ pour s. Le passage de s à θ, grâce à $\hat{\theta}(s)$, est l'inférence statistique, dont les propriétés ne sont connues que si le choix de s est aléatoire.

Un plan de sondage est une loi de probabilité sur l'ensemble $\mathscr{P}(U)$ des parties (échantillons) de U. La notation classique d'un échantillon aléatoire à valeur dans $\mathscr{P}(U)$ est $\mathbb{S}$. Un plan de sondage pour lequel tous les échantillons de taille différente de n ($n \in \mathbb{N}^*$) ont une probabilité nulle d'être sélectionnés est dit de taille fixe n. La manipulation d'une loi de probabilité sur $\mathscr{P}(U)$ est généralement complexe. C'est pourquoi le statisticien d'enquête travaille avec des résumés de la loi de $\mathbb{S}$: les probabilités d'inclusion simple et double. Elles font respectivement référence aux probabilités qu'un individu donné ou qu'un couple donné d'individus soit sélectionné : $\pi_i = \mathbb{P}(i \in \mathbb{S})$ et $\pi_{ij} = \mathbb{P}((i,j) \in \mathbb{S})$.

1. Les unités primaires constituent une partition de la population selon des critères géographiques. La sélection dans un premier temps d'UP puis d'individus dans ces UP est de nature à concentrer la collecte et à réduire les coûts lorsque l'enquête se déroule en face-à-face.

2. Dans ce chapitre, contrairement aux chapitres précédents, la notion de "taille" d'une zone géographique renvoie au nombre d'entités présentes à l'intérieur, et non pas à sa surface.

L'estimation de θ par $\hat{\theta}$ est entachée d'erreurs multiples :
— erreur d'échantillonnage
— erreur de couverture : existence d'individus ne pouvant jamais être sélectionnés ;
— erreur de non-réponse : existence d'individus pour lesquels la valeur de y_i est inconnue alors même qu'ils sont échantillonnés ;
— erreur de mesure : fait de récolter la valeur y_i^* au lieu de y_i.

Un estimateur dont l'espérance est différente de θ est biaisé, alors que la variabilité des différentes valeurs $\hat{\theta}(s)$ est appréhendée par la variance de $\hat{\theta}$. L'objectif est de rendre le biais et la variance aussi petits que possible, en prêtant une attention particulière aux conditions de collecte de l'information et/ou en choisissant judicieusement le plan de sondage.

Parmi les différents paramètres à estimer, le plus classique est le total d'une variable d'intérêt : $\theta = t_y = \sum_{i \in U} y_i$. Parmi les différents estimateurs possibles de t_y, on s'intéresse à l'estimateur de Narain-Horvitz-Thompson : $\hat{t}_y = \sum_{i \in \mathbb{S}} y_i/\pi_i$. En l'absence d'erreurs de couverture, de non-réponse, et de mesure, cet estimateur est sans biais. Sa variance pour un plan de taille fixe est :

$$V(\hat{t}_y) = -\frac{1}{2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{10.1}$$

L'analyse de l'équation 10.1 donne des indications quant aux plans de sondage à retenir pour estimer de manière précise la quantité t_y. Si les π_i sont proportionnelles aux y_i, la variance est nulle. Par rapport à cet idéal inatteignable, l'alternative de second rang consisterait à retenir, dans le cas d'une enquête, des π_i proportionnelles à x_i, où x est une variable auxiliaire connue pour tous les individus de U et qui est corrélée à y.

Ce résultat est valable quand l'enquête est mono-thème (une seule variable d'intérêt y). L'utilisation de telles probabilités pour une autre variable d'intérêt y' non corrélée à x peut en effet conduire à des estimations très imprécises. C'est pourquoi, quand l'enquête est multi-thèmes, le statisticien préfère souvent choisir des probabilités d'inclusion simple constantes. Elle permet "de rendre minimales les variances que l'on obtiendrait dans les configurations les plus défavorables (on parle d'optique MINIMAX), [...] c'est-à-dire pour les variables qui sont le plus de nature à détériorer la précision des estimations" (ARDILLY 2006). À probabilités d'inclusion simple fixées, il est souhaitable que le plan attribue des π_{ij} grands lorsque l'écart entre y_i/π_i et y_j/π_j est grand, et vice versa. Dans le cas de variables spatialisées, si on suppose que deux individus proches géographiquement ont des valeurs proches de y_i/π_i et y_j/π_j, il conviendra, à probabilités d'inclusion fixées, de privilégier la sélection d'individus éloignés plutôt que proches.

10.2 Constituer des unités primaires de faible étendue et de taille constante

10.2.1 Pourquoi ?

Quand les contraintes organisationnelles se traduisent par un mode de collecte en face-à-face sur un territoire dont la densité de population est faible, la méthode généralement retenue est celle du sondage à deux degrés. Afin de réduire les coûts liés aux déplacements des enquêteurs, le premier degré conduit à la sélection d'entités géographiques (unités primaires, UP) dont l'étendue géographique doit être la plus petite possible. De manière simplifiée, une UP ainsi sélectionnée est affectée ensuite à un seul enquêteur. Au sein de chaque UP sont sélectionnées des unités secondaires (US) correspondant aux unités statistiques que l'on souhaite interroger (individus dans des résidences principales, entreprises). Afin d'assurer une charge de travail suffisante à l'enquêteur sur une ou plusieurs enquêtes, chaque UP doit par ailleurs comporter un nombre minimal d'unités secondaires.

Pour un réseau constitué de m enquêteurs et un échantillon final de n unités secondaires, m UP sont sélectionnées proportionnellement à leur nombre d'unités secondaires : $\pi_i^{(1)} = m(N_i/N)$ pour l'UP i regroupant un total de N_i unités secondaires. En admettant que m divise n, dans chacune de ces m UP, n/m US sont tirées selon un plan à probabilités égales : $\pi_j^{i(2)} = n/(mN_i)$ pour l'unité secondaire j de l'UP i. La probabilité d'inclusion finale est constante : $\pi_j^i = \pi_i^{(1)}\pi_j^{i(2)} = n/N$.

Quand la maille géographique la plus fine disponible dans la base de sondage reste grossière, par exemple communale, les UP sont constituées par regroupement de ces mailles. Il est plus délicat d'en maîtriser la taille. Le plan ne bénéficie pas, au premier degré de tirage, de la propriété MINIMAX évoquée précédemment puisque les probabilités sont proportionnelles à la taille. Il est possible de retrouver cette propriété si les UP sont de taille constante. La constance de la taille des UP peut par ailleurs être souhaitable pour d'autres caractéristiques liées à la coordination des échantillons (sélection d'échantillons disjoints ou d'échantillons inclus les uns dans les autres). On note que :

— le complémentaire $\overline{\mathbb{S}} = U \backslash \mathbb{S}$ d'un échantillon aléatoire $\mathbb{S}$ tiré à probabilités égales, est lui-même à probabilités égales ;
— un échantillon aléatoire $\mathbb{S}_2$, sélectionné à probabilités constantes dans un échantillon $\mathbb{S}_1$ lui-même sélectionné à probabilités constantes, est à probabilités constantes.

L'idéal est donc de construire des UP de faible étendue géographique et ayant le même nombre d'unités secondaires.

10.2.2 Comment ?

Le problème de constitution d'unités primaires de faible étendue géographique et de taille constante est un cas particulier du problème plus général de classification sous contraintes de taille, qui a connu un intérêt renouvelé dans la littérature récente (MALINEN et al. 2014, GANGANATH et al. 2014, TAI et al. 2017). Il s'agit en effet d'obtenir une partition du territoire en classes, à l'intérieur desquelles la dispersion des coordonnées géographiques est la plus faible possible tout en ayant un nombre donné d'unités par classe. On présente ici une méthode mise en place initialement pour la constitution d'UP dans le cadre de l'enquête Emploi française (LOONIS 2009), et reprise récemment dans le cadre de travaux pour la constitution d'UP de l'échantillon maître français (ensemble des enquêtes auprès des ménages) (FAVRE-MARTINOZ et al. 2017).

Le principe général est le suivant :

1. On considère le géoréférencement le plus fin possible des unités statistiques. Du fait de la qualité du géoréférencement ou de la nature des données, le nombre n_{xy} d'unités statistiques localisées au point de coordonnées $(x;y)$ peut être strictement supérieur à 1.
2. On fait passer un chemin parmi l'ensemble des localisations connues. On utilise pour cela les méthodes évoquées dans le chapitre 2 : "Codifier la structure de voisinage". Dans la mesure où il n'y a pas de nécessité que le chemin revienne à son point de départ, on privilégie le chemin de Hamilton, qui est le plus court (ce chemin minimise la somme des distances entre deux points consécutifs sans fixer de point de départ ou d'arrivée).
3. Pour construire M zones, on parcourt le chemin depuis son origine en cumulant les quantités n_{xy}. Quand ce total dépasse le seuil $c \simeq \frac{N}{M}$, la première UP est constituée. On reprend alors le processus à partir du premier point non encore visité du chemin.
4. Dans des conditions idéales où c divise N et $n_{xy} = 1$ pour tout couple (x,y), la procédure conduit à des unités primaires homogènes géographiquement et de taille constante. Cette heuristique ne conduit cependant pas à un optimum global. Il convient de prévoir, comme dans toute classification, une procédure de consolidation permettant de gérer les éventuelles situations géographiques atypiques et/ou les tailles d'UP trop éloignées de c. Cette dernière situation peut se présenter, par exemple, quand le dernier point du chemin intégré dans une

UP correspond à une valeur très élevée de n_{xy}, ou pour la dernière UP constituée.

Dans la partie qui suit, on montre une application de cette procédure en se focalisant plus particulièrement sur la question de la construction du chemin quand le nombre d'unités secondaires est important.

10.2.3 Application

La figure 10.1 montre les résultats d'une application de la stratégie générale précédente à la région Alsace (ancienne région, avant la restructuration des régions françaises en 2016). Pour les besoins de l'enquête Emploi en continu (EEC), les résidences principales de la région sont regroupées en unités primaires de 2 600 résidences principales (figure 10.1b), elles-mêmes découpées en secteurs ayant chacun 120 résidences principales (figure 10.1c).

Pour des raisons de temps de calcul dans la constitution des UP, les quelques 616 000 résidences principales ont été initialement regroupées dans 80 000 carreaux de 100 mètres de côté (figure 10.1a), qui constituent donc le géoréférencement le plus fin des unités statistiques. Pour la construction des secteurs au sein des UP, les résidences principales sont par nature géocodées à l'immeuble. La variabilité de la taille des carreaux d'origine ou des immeubles, et donc des n_{xy}, explique en partie la légère variabilité de la taille des UP et de la taille des secteurs finalement obtenues (tableau 10.1).

Ordre du fractile	Taille du carreau d'origine	Taille de l'UP (figure 10.1b)	Taille du secteur (figure 10.1c)
100 %	378	2776	139
99 %	59	2757	131
95 %	23	2685	130
90 %	15	2640	130
75 %	9	2606	124
50 %	5	2595	119
25 %	2	2591	118
10 %	1	2587	118
5 %	1	2502	117
1 %	1	2491	111
0 %	1	2479	99

TABLE 10.1 – Fractiles en nombre de résidences principales dans les carreaux, les UP et les secteurs de la figure 10.1

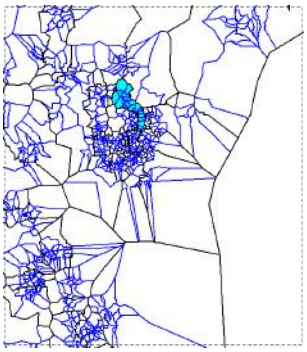
La constitution des secteurs de taille 120 à partir d'un grand nombre de résidences principales peut vite poser des problèmes de temps de calcul. Avec une distance euclidienne, la constitution du chemin de Hamilton le plus court peut être exacte dès lors que le nombre de points n'excède pas quelques centaines. Quand il y en a plusieurs milliers, il n'est pas plus raisonnable qu'utile de chercher à construire le chemin optimal de façon exacte. Nous proposons donc une méthode approchée visant à construire rapidement un chemin qui aura de bonnes propriétés compte tenu de l'objectif fixé. Les différentes étapes de cette méthode approchée sont décrites ci-dessous et illustrées en figure 10.2 pour une UP prise en exemple. Celle-ci comporte 2 600 résidences principales et 1 085 immeubles.

1. Les 1 085 immeubles et 2 600 résidences principales de l'UP en bleu dans la figure 10.1c, sont regroupés par la méthode des **nuées dynamiques** (ou *k-means*) [3] en 20 classes d'effectifs différents mais cohérentes géographiquement. Les variables sur lesquelles s'effectue cette

3. La méthode *k-means* vise à créer des classes homogènes en maximisant la variance entre les classes et en minimisant la variance au sein de chacune d'entre elles.

(a) 616 000 résidences principales, dans 80 000 carreaux de 100 m de côté ...

(b) ... sont réparties dans des UP homogènes de 2 600 résidences principales ...



(c) ... et découpées en secteurs de 120 résidences principales.

FIGURE 10.1 – Construction de zones de faible étendue géographique et de taille constante en nombre de résidences principales en Alsace

classification sont les coordonnées géographiques des immeubles. On notera que $20 \simeq \frac{2600}{120}$ (figures 10.2a et 10.2b).

2. On fait passer un **chemin** de Hamilton parmi les barycentres de ces 20 classes de manière à pouvoir les ordonner (figure 10.2c).
3. Dans une classe donnée i, on **ordonne** les immeubles selon **deux sous-parties** (figure 10.2d) :
   (a) la première regroupe les immeubles de la classe i qui sont plus près de G_{i-1} (barycentre de la classe précédente) que de G_{i+1} (barycentre de la classe suivante), triés par distance croissante à G_{i-1} ;
   (b) la seconde regroupe les immeubles de la classe i qui sont plus près G_{i+1} (barycentre de la classe suivante) que de G_{i-1} (barycentre de la classe précédente), triés par distance décroissante à G_{i+1}.
4. Par construction, les premiers immeubles de la classe i sont proches des derniers immeubles de $i-1$, et les derniers de i sont proches des premiers de $i+1$. Le chemin revient ainsi à parcourir les immeubles par classe, par sous-partie puis par distance croissante ou décroissante selon le cas (figure 10.2e). À l'intérieur d'un immeuble, si besoin, il est possible de trier les logements par étage.

10.3 Comment sélectionner un échantillon spatialement dispersé ?

Les considérations générales ont montré qu'une estimation sera d'autant plus précise que le plan de sondage privilégie la sélection d'individus géographiquement éloignés les uns des autres. GRAFSTRÖM et al. 2013, par exemple, ont formalisé ces considérations de manière plus explicite. Dans cette partie, nous détaillons les méthodes permettant de sélectionner des échantillons spatialement dispersés. Ces méthodes peuvent être regroupées en deux familles.

Pour les méthodes de la première famille, les probabilités d'inclusion des unités sont mises à jour localement afin de limiter la sélection de deux unités voisines. Dans cette famille, on retrouve la méthode du tirage poissonien corrélé spatialement (GRAFSTRÖM 2012), la méthode du pivot spatial (GRAFSTRÖM et al. 2012), et la méthode du cube spatial (GRAFSTRÖM et al. 2013). La deuxième famille se caractérise par une transformation du problème de proximité des unités dans plusieurs dimensions en un problème d'ordre dans $\mathbb{R}$. Ensuite, le tirage s'effectue selon un tirage excluant deux unités proches dans le fichier trié. Cette famille de méthodes regroupe la méthode *general randomized tessellation stratified* (GRTS, STEVENS JR et al. 2004), la méthode basée sur une courbe de Peano (LISTER et al. 2009) ou sur l'algorithme du voyageur de commerce (DICKSON et al. 2016).

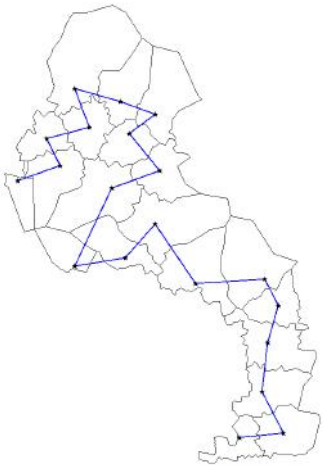
10.3.1 Le tirage poissonien corrélé spatialement (GRAFSTRÖM 2012)

Le tirage poissonien corrélé spatialement est une extension du tirage poissonien corrélé (*Correlated Poisson Sampling*, CPS) proposé par BONDESSON et al. 2008 pour réaliser de l'échantillonnage en temps réel. La méthode CPS est fondée sur un échantillonnage séquentiel et ordonné des unités. Les unités sont ordonnées avec des indices allant de 1 à N. On statue d'abord sur l'unité 1, puis sur l'unité 2, jusqu'à l'unité N. Dans le cas de l'échantillonnage en temps réel, l'ordre de l'indiçage correspond à un ordre de visite préétabli des unités échantillonnables. Dans le cas d'un tirage spatial, l'ordre peut être établi à partir de la proximité des unités selon une fonction de distance euclidienne. À chaque étape, les probabilités d'inclusion sont mises à jour de façon à engendrer une corrélation positive ou négative entre les indicatrices de sélection des unités.

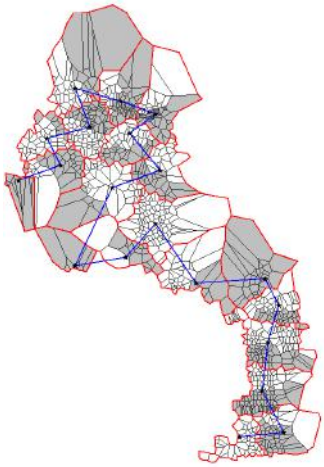
Plus précisément, la première unité est incluse dans l'échantillon avec une probabilité $\pi_1^0 = \pi_1$. On note I_1 l'indicatrice qui vaut 1 si cette unité est sélectionnée, 0 sinon. Plus généralement, à l'étape j, on sélectionne l'unité j avec la probabilité π_j^{j-1} et on met à jour les probabilités

(a) Les immeubles, et leurs cellules de Voronoï associées ...

(b) ... sont regroupés, par nuées dynamiques sur les coordonnées, en une vingtaine de classes de taille variable.



(c) Un chemin passant par les barycentres des classes ...



(d) ... permet de classer les immeubles selon qu'ils sont plus proches du barycentre de la classe précédente (blanc) ou de la classe suivante (gris) ...

(e) ... et donc de créer un chemin passant par tous les immeubles.

(f) En suivant ce chemin, on construit des secteurs de 123 à 128 résidences principales.

FIGURE 10.2 – Des logements aux secteurs de 120 logements

d'inclusion des unités $i \geq j+1$ de la façon suivante :

$$\pi_i^j = \pi_i^{j-1} - (I_j - \pi_j^{j-1}) w_j^i, \tag{10.2}$$

où les w_j^i sont les poids donnés par l'unité j aux unités avec des indices $i \geq j+1$. Les probabilités d'inclusion sont mises à jour étape par étape, avec au plus N étapes jusqu'à l'obtention du vecteur des indicatrices de sélection.

Le choix des poids w_j^i est crucial, car il permet de déterminer si l'on introduit une corrélation positive ou négative entre les indicatrices de sélection. BONDESSON et al. 2008 donnent l'expression de ces poids pour quelques plans de sondage classiques, et une expression générale pour tout plan de sondage. Ainsi, cette méthode est très générale : tout plan de sondage pour lequel les probabilités d'inclusion simple sont fixées peut être implémenté par la méthode CPS. Seules l'expression et les conditions qui portent sur les poids [4] vont varier d'un plan à un autre : par exemple, pour un plan de taille fixe, la somme des poids $w_j^i,\ (j<i)$ doit être égale à 1. Dans un contexte d'autocorrélation spatiale positive (où les unités proches sont semblables), les poids associés doivent être choisis positivement, de façon à introduire une corrélation négative entre les indicatrices de sélection. Il semble donc pertinent de réaliser un test d'autocorrélation spatiale globale pour déterminer le signe des poids à mobiliser dans cette méthode.

GRAFSTRÖM 2012 propose dans son article deux versions pour les poids w_j^i. Nous présentons ici la version considérant une distribution gaussienne. Dans ce cas, les poids sont de la forme :

$$w_j^i \propto \exp(-[d(i,j)/\sigma]^2),\ i = j+1, j+2, ..., N \tag{10.3}$$

.

La constante de proportionnalité est fixée par le fait que la somme des poids doit être égale à 1. Ces poids sont d'autant plus grands que les unités sont proches de l'unité j. Ainsi, la probabilité π_i^j sera d'autant plus faible que l'unité i sera proche (au sens de la distance $d(i,j)$) de l'unité j, ce qui permet de réaliser un échantillonnage spatialement dispersé. Le paramètre σ permet de gérer la dispersion de ces poids, et donc de répartir la mise à jour des probabilités d'inclusion simple dans un voisinage plus ou moins large.

Cette méthode est implémentée sous R dans le package *BalancedSampling* (GRAFSTRÖM et al. 2016) avec la fonction `scps()`.

10.3.2 La méthode du pivot spatial (GRAFSTRÖM et al. 2012)

Rappel sur la méthode du pivot

La méthode du pivot est une procédure d'échantillonnage permettant de sélectionner un échantillon avec des probabilités d'inclusion égales ou inégales (DEVILLE et al. 1998). À chaque étape de l'algorithme, les probabilités d'inclusion de deux unités i et j en lice sont mises à jour, et l'une au moins de ces deux unités est sélectionnée ou rejetée. Le vecteur des probabilités d'inclusion des deux unités en lice (π_i, π_j) est mis à jour selon la règle suivante (combat entre les unités i et j) :

— si $\pi_i + \pi_j < 1$, alors :

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ avec la probabilité } \dfrac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ avec la probabilité } \dfrac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

4. les conditions imposées au poids sont liées aux conditions imposées aux probabilités d'inclusion simple, donc au plan de sondage.

— si $\pi_i + \pi_j \geq 1$, alors :

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ avec la probabilité } \dfrac{(1-\pi_j)}{(2-\pi_i-\pi_j)} \\ (\pi_i + \pi_j - 1, 1) \text{ avec la probabilité } \dfrac{(1-\pi_i)}{(2-\pi_i-\pi_j)} \end{cases}$$

Cette procédure est répétée jusqu'à l'obtention d'un vecteur des probabilités d'inclusion contenant $N-n$ fois le chiffre 0 et n fois le chiffre 1, déterminant complètement l'échantillon sélectionné (au plus N étapes).

Extension à l'échantillonnage spatial

La méthode du pivot spatial (GRAFSTRÖM et al. 2012) est une extension spatiale de la méthode du pivot. L'idée de la méthode est toujours de mettre à jour le vecteur des probabilités d'inclusion π de façon successive, mais en sélectionnant cette fois-ci à chaque étape deux unités voisines au sens d'une certaine distance (e.g. une distance euclidienne) pour participer au combat. Plusieurs méthodes permettent de sélectionner ces deux unités voisines :

— **LPM1** : deux unités les plus proches l'une de l'autre sont sélectionnées pour participer au combat, *i.e.* une unité i est sélectionnée aléatoirement parmi les N unités de la population, puis l'unité j la plus proche de i est sélectionnée pour participer au combat si et seulement si i est aussi l'unité la plus proche de j (au mieux N^2 étapes, au pire N^3 étapes) ;
— **LPM2** : deux unités voisines sont sélectionnées pour participer au combat, *i.e.* une unité i est sélectionnée aléatoirement parmi les N unités de la population, puis l'unité j la plus proche de i est sélectionnée pour participer au combat (N^2 étapes) ;
— **LPM K-D TREE** : les deux unités voisines sont sélectionnées au moyen d'un arbre *k-d* de partition de l'espace (LISIC 2015) permettant de faire les recherches de plus proches voisins plus rapidement (complexité de l'algorithme en $N\log(N)$).

Ces trois méthodes du pivot spatial sont implémentées en C++ dans le package *BalancedSampling* du logiciel R.

10.3.3 La méthode du cube

Généralités sur la méthode du cube

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

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

$$\sum_{i \in \mathbb{S}} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i \tag{10.4}$$

Un algorithme permettant de réaliser un tel tirage est l'algorithme du cube. Pour en décrire le principe, il est opportun d'avoir recours à la représentation géométrique suivante. Un échantillon est un des sommets d'un hypercube de dimension N, noté C. L'ensemble des p contraintes, rappelées par l'équation (10.4), définit un hyperplan de dimension $N-p$, noté Q. On note $K = Q \cap C$, l'intersection du cube et de l'hyperplan. Une représentation graphique du problème en dimension 3, tirée de l'article DEVILLE et al. 2004, est donnée ci-dessous (figure 10.3).

L'algorithme du cube se décompose en deux phases. La première phase, dite "phase de vol" (figure 10.4), est une marche aléatoire qui part du vecteur des probabilités d'inclusion et évolue dans K. Pour cela, on part de $\pi(0) = \pi$, puis on met à jour le vecteur des probabilités d'inclusion

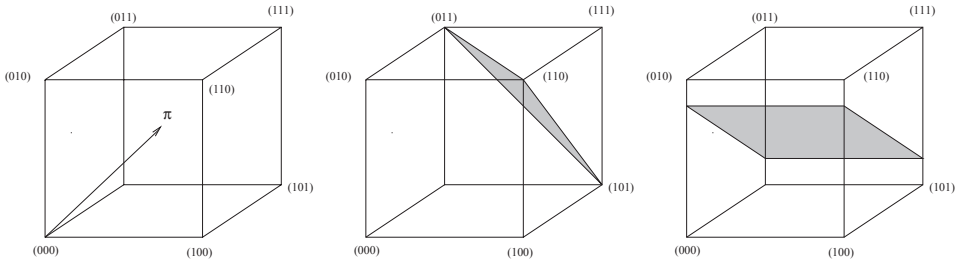


FIGURE 10.3 – Représentation graphique du cube pour $N = 3$ et différentes configurations possibles de l'espace des contraintes avec ici $p = 1$

en choisissant un vecteur $u(0)$ de sorte que $\pi + u(0)$ demeure dans l'espace des contraintes. En suivant la direction donnée par le vecteur $u(0)$, on aboutit nécessairement sur une face du cube. Le sens pour la mise à jour du vecteur des probabilités d'inclusion est ensuite donné par les paramètres $\lambda_1^*(0)$ et $\lambda_2^*(0)$, ceux-ci étant choisis de sorte que le vecteur mis à jour $\pi(1)$ touche une face du cube. Le choix du sens pour cette mise à jour est effectué de façon aléatoire de façon que $E(\pi(1)) = \pi(0)$. On recommence ensuite l'opération en choisissant un nouveau vecteur $u(1)$ pour la direction et un nouveau sens pour la mise à jour des probabilités d'inclusion. Cette marche aléatoire s'arrête lorsqu'elle a atteint un sommet π^* de K. À l'issue de cette première phase, le sommet π^* n'est pas nécessairement un sommet du cube C. Soit q, le nombre de composantes non entières dans le vecteur π^* ($q \leq p$). Si q est nul, la procédure d'échantillonnage est terminée, sinon il faut procéder à la deuxième étape, appelée "phase d'atterrissage". Elle consiste à relâcher le moins possible les contraintes d'équilibrage, et à relancer une phase de vol avec ces nouvelles contraintes jusqu'à l'obtention d'un échantillon. Il n'est pas envisageable de modifier dès le début l'espace des contraintes de sorte que les sommets de K soient confondus avec ceux de C, car cela reviendrait à tester tous les échantillons possibles pour voir dans un premier temps si l'un d'eux permet de respecter les contraintes. Le fait de modifier l'espace des contraintes dans un deuxième temps (dans la phase d'atterrissage) permet de travailler sur une population U^* de plus petite taille ($\dim(U^*) = q$). Le problème peut ainsi être résolu car le nombre d'échantillons à considérer est raisonnable.

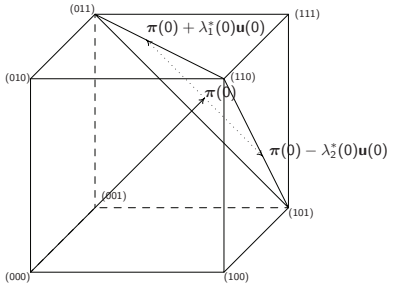


FIGURE 10.4 – Première étape de la phase de vol du cube pour $N = 3$ et une contrainte ($p = 1$) d'un échantillon de taille fixe $n = 2$

L'implémentation de cet algorithme est disponible sous SAS grâce à la macro *FAST CUBE* ou sous R dans le package *BalancedSampling*.

La méthode du cube spatial

L'idée générale de l'algorithme de tirage spatialement équilibré est de construire un cluster de $p+1$ unités géographiquement proches, puis d'appliquer la phase de vol du cube sur ce cluster. Cela

conduit à statuer sur la sélection ou non d'une unité dans ce cluster en respectant les p contraintes localement dans ce cluster. Ensuite, les probabilités sont modifiées localement, ce qui assure que les probabilités d'inclusion des unités proches sont réduites si l'unité sur laquelle on a statué est sélectionnée. Cela limite ainsi la probabilité qu'une de ses unités proches soit sélectionnée dans l'étape suivante de l'algorithme. Puis on répète la procédure : on sélectionne une unité, on crée un cluster de $p+1$ unités autour de l'unité sélectionnée, et on applique la phase de vol du cube avec les probabilités d'inclusion mises à jour à l'étape précédente. On répète le processus tant qu'il reste plus de $p+1$ unités. Pour finir, on applique la phase d'atterrissage classique du cube.

La méthode de tirage spatialement équilibrée décrite ci-dessus est disponible dans le package R *BalancedSampling*. Ce package, développé en C++, permet d'appliquer l'algorithme très rapidement.

Équilibrage sur les moments

La définition d'un échantillon spatialement équilibré laisse entrevoir une utilisation différente de l'algorithme du cube pour l'échantillonnage spatial. Pour MARKER et al. 2009, "un échantillon est spatialement équilibré si les moments spatiaux des échantillons localisés correspondent aux moments spatiaux de la population. Les moments spatiaux sont le centre de gravité et l'inertie.". Dans la terminologie de l'algorithme du cube, cette définition peut revenir à sélectionner un échantillon équilibré sur des variables définies à partir des coordonnées géographiques : $x_i, y_i, x_i^2, y_i^2, x_i y_i$ afin de respecter les moments non centrés d'ordre 1 et 2 :

— $T_x = \sum_{i \in U} x_i$,
— $T_y = \sum_{i \in U} y_i$,
— $T_{x^2} = \sum_{i \in U} x_i^2$,
— $T_{y^2} = \sum_{i \in U} y_i^2$,
— $T_{xy} = \sum_{i \in U} x_i y_i$.

10.3.4 Méthodes sur fichier trié

Les méthodes existantes dans cette famille (STEVENS JR et al. 2004, DICKSON et al. 2016, LISTER et al. 2009) s'appuient dans un premier temps sur la constitution d'un chemin passant par toutes les unités statistiques. Ce chemin peut être GRTS (*generalized random tessellation stratified*), voyageur de commerce (TSP), ou une courbe de Peano. Conditionnellement à l'ordre défini par ce chemin, il s'agit ensuite de sélectionner un échantillon selon une méthode qui exclut deux unités proches, par exemple le tirage systématique.

D'autres méthodes de constitution de chemin existent (chemins de Hamilton, ou courbes remplissant l'espace : Hilbert, Lebesgue). De même, il existe d'autres méthodes de sélection excluant les unités proches à tri donné, comme celle des plans de sondage déterminantaux (LOONIS et al. 2018). La question des chemins ayant été abordée dans le chapitre 2 : "Codifier la structure de voisinage", on présente ici les propriétés répulsives des plans systématiques et déterminantaux.

La méthode de tirage systématique

Le tirage systématique est une méthode de tirage simple à mettre en œuvre et qui permet de réaliser un tirage à probabilités inégales tout en respectant les probabilités d'inclusion simple. Cette méthode a été proposée par MADOW 1949, puis étendue par CONNOR 1966, BREWER 1963, PINCIARO 1978, et HIDIROGLOU et al. 1980. Il est très souvent utilisé en pratique pour les enquêtes par téléphone, pour faire de l'échantillonnage sur des flux continus de données, ou dans le tirage des logements dans le cas des enquêtes ménages de l'Insee.

Pour tirer un échantillon de taille fixe n respectant le vecteur de probabilités d'inclusion π, on commence par définir la somme cumulée des probabilités d'inclusion par $V_i = \sum_{l=1}^{i} \pi_l, i \in U$, avec $V_0 = 0$. Pour un échantillon de taille fixe, on a $V_N = n$. Ensuite, on utilise l'algorithme du tirage systématique présenté ci-dessous pour statuer sur les unités à échantillonner.

Algorithme du tirage systématique :

— Générer une variable aléatoire u uniformément distribuée sur l'intervalle $[0,1]$.

— Pour $i = 1,...,N$,

$$I_i = \begin{cases} 1 & \text{si il existe un entier } j \text{ tel que } V_{i-1} \leq u + j - 1 < V_i, \\ 0 & \text{sinon.} \end{cases}$$

Le tableau 10.2 présente un exemple de la méthode dans le cas où $n = 3$ et $N = 10$.

i	1	2	3	4	5	6	7	8	9	10
π_i	0.2	0.2	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.3
V_i	0.2	0.4	0.7	1	1.4	1.8	2.1	2.4	2.7	3(=n)

TABLE 10.2 – Un exemple de tirage systématique

Par exemple si le nombre aléatoire généré u est égal à 0.53, les unités 2, 5, 8 seront sélectionnées car satisfaisant les contraintes :

$$V_2 \leq u < V_3 \qquad V_5 \leq u+1 < V_6 \qquad V_8 \leq u+2 < V_9.$$

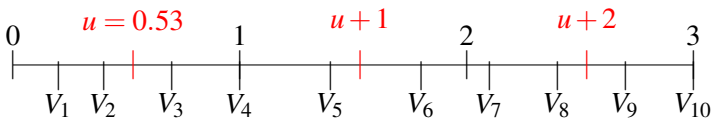


FIGURE 10.5 – Sélection de 3 unités parmi 10

Selon cette méthode, les unités (i, j) telles que $|V_i - V_j| < 1$ ont une probabilité nulle d'être sélectionnées ensemble. Si le fichier est trié judicieusement, cela assure la dispersion géographique de l'échantillon.

Implémentation de la méthode GRTS

Parmi les pratiques de tirage systématique sur fichier trié géographiquement, la méthode GRTS est très utilisée. Le tri GRTS est décrit dans le chapitre 2 : "Codifier la structure géographique". Le package *gstat* du logiciel R a été implémenté spécifiquement pour tirer des échantillons à l'aide de cette méthode. Cependant, la méthode GRTS présente quelques désavantages, en particulier le fait que l'algorithme de découpage et celui de tirage ne soient pas dissociés, ou encore le temps d'exécution de la méthode. En effet, la méthode propose par défaut de s'arrêter à 11 niveaux hiérarchiques pour le découpage, le temps d'exécution de la méthode risquant d'être trop important si un découpage plus fin était demandé. Au-delà, si plus d'une unité appartiennent à la même cellule, un échantillon est sélectionné au sein de la cellule à l'aide de la fontion *S-PLUS*, prenant pour paramètre *prob* qui correspond aux probabilités d'inclusion de chaque élément. L'algorithme de GRTS s'adapte ainsi difficilement à des populations de grande taille. Afin de pallier ces limites d'ordre computationnel, une nouvelle méthode du Pivot par Tessellation (CHAUVET et al. 2017) a

été développée en R, où l'algorithme de Tessellation (très proche de celui de GRTS) est dissocié de la partie tirage. Cette méthode s'appuie sur la décomposition binaire d'un nombre, ce qui permet d'effectuer le découpage directement sur 31 niveaux. Le temps d'exécution est donc considérablement amélioré. De plus, il est possible d'effectuer ce découpage dans plus de deux dimensions.

Échantillonnage déterminantal

Par définition, une variable aléatoire $\mathbb{S}$ à valeurs dans 2^U a pour loi de probabilité un plan de sondage déterminantal si et seulement si il existe une matrice hermitienne contractante [5] K indexée par U, appelée noyau, telle que pour tout $s \in 2^U$,

$$p(s \subseteq \mathbb{S}) = \det(K_{|s}) \tag{10.5}$$

où $K_{|s}$ est la sous matrice de K indexée par les unités de s. De cette définition découle directement le calcul des probabilités d'inclusion (table 10.3).

π_i	=	$pr(i \in \mathbb{S})$	=	$\det(K_{\|\{i\}})$	=	K_{ii}
π_{ij}	=	$pr(i,j \in \mathbb{S})$	=	$\det\begin{pmatrix} K_{ii} & K_{ij} \\ \overline{K}_{ij} & K_{jj} \end{pmatrix}$	=	$K_{ii}K_{jj} - \mid K_{ij} \mid^2$

TABLE 10.3 – Calcul des probabilités d'inclusion simple et double dans un plan de sondage déterminantal de noyau K
Note : $|z|$ désigne le module du nombre complexe z.

La diagonale de la matrice K correspond aux probabilités d'inclusion simple. Un autre résultat particulièrement important des plans déterminantaux est celui précisant qu'un plan déterminantal est de taille fixe si et seulement si K est une matrice de projection [6](HOUGH et al. 2006).

On considère l'ensemble des matrices de projection dont la diagonale correspond à un vecteur Π de probabilités d'inclusion fixées *a priori*. Parmi elles, la matrice K^Π (dont les coefficients sont donnés par la table 10.4) présente des propriétés intéressantes en termes de répulsion spatiale.

	Valeurs de j	
Valeurs de i	$j = i_r$	$i_r < j < i_{r+1}$
$i_{r'} < i < i_{r'+1}$	$-\sqrt{\Pi_i}\sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_l)}{1-(\Pi_j-\alpha_j)}}\gamma_r^{r'}$	$\sqrt{\Pi_i\Pi_j}\gamma_r^{r'}$
$i = i_{r'+1}$	$-\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}}\sqrt{\frac{(1-\Pi_j)(\Pi_j-\alpha_j)}{1-(\Pi_j-\alpha_j)}}\gamma_r^{r'}$	$\sqrt{\frac{(1-\Pi_i)\alpha_i}{1-\alpha_i}}\sqrt{\Pi_j}\gamma_r^{r'}$

où pour tout r tel que $1 \leq r \leq n$:

— $1 < i_r \leq N$ est un entier tel que $\sum_{i=1}^{i_r-1} \Pi_i < r$ et $\sum_{i=1}^{i_r} \Pi_i \geq r$; par convention on posera $i_0 = 0$

— $\alpha_{i_r} = r - \sum_{i=1}^{i_r-1} \Pi_i$. On notera que $\alpha_{i_r} = \Pi_{i_r}$ si $\sum_{i=1}^{i_r} \Pi_i = r$.

— $\gamma_r^{r'} = \sqrt{\prod_{k=r+1}^{r'} \frac{(\Pi_{i_k}-\alpha_{i_k})\alpha_{i_k}}{(1-\alpha_{i_k})(1-(\Pi_{i_k}-\alpha_{i_k}))}}$ pour $r < r'$, $\gamma_r^{r'} = 1$ autrement.

TABLE 10.4 – Valeurs de K_{ij}^Π avec $i > j$

5. Une matrice complexe K est hermitienne si $K = \overline{K}^t$, où les coefficients de $\overline{K}$ sont les conjugués de ceux de K. Une matrice est contractante si toutes ses valeurs propres sont comprises entre 0 et 1.

6. Une matrice hermitienne est de projection si ses valeurs propres sont 0 ou 1.

La répulsion du plan déterminantal associé à K^{Π} pour des individus proches (selon l'ordre du fichier) est illustrée par les propriétés suivantes (LOONIS et al. 2018) :

1. le plan sélectionnera au plus un individu dans un intervalle de la forme $]i_r+1, i_{r+1}-1[$;
2. si un individu y est tiré, ainsi que l'individu "proche" i_{r+1}, alors le plan ne sélectionnera pas d'individu supplémentaire "proche", c'est-à-dire dans $]i_{r+1}+1, i_{r+2}-1[$;
3. ce plan aura toujours au moins un individu dans un intervalle $[i_r+1, i_{r+1}-1]$;
4. si $|i-j|$ est grand, alors $\pi_{ij} \approx \Pi_i \Pi_j$. On retrouve les probabilités d'inclusion double du plan poissonien.

L'application des résultats sur les plans déterminantaux aux probabilités définies dans la table 10.2 conduit aux quantités : $i_1 = 4, i_2 = 7, i_3 = 10$ et $\alpha_4 = 0.3 = \Pi_4, \alpha_7 = 0.2, \alpha_{10} = 0.3 = \Pi_{10}$. Les probabilités d'inclusion doubles sont données par la matrice ci-dessous.

$$\begin{pmatrix}
 & 0 & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\
0 & & 0 & 0 & \frac{2}{25} & \frac{2}{25} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} & \frac{3}{50} \\
0 & 0 & & 0 & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} \\
0 & 0 & 0 & & \frac{3}{25} & \frac{3}{25} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} & \frac{9}{100} \\
\frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & & 0 & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\
\frac{2}{25} & \frac{2}{25} & \frac{3}{25} & \frac{3}{25} & 0 & & \frac{1}{20} & \frac{7}{60} & \frac{7}{60} & \frac{7}{60} \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{1}{20} & \frac{1}{20} & & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & & 0 & 0 \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & 0 & & 0 \\
\frac{3}{50} & \frac{3}{50} & \frac{9}{100} & \frac{9}{100} & \frac{7}{60} & \frac{7}{60} & \frac{1}{15} & 0 & 0 &
\end{pmatrix}.$$

Les termes autour de la diagonale principale ont tendance à être nuls ou proches de 0, traduisant la répulsion.

10.4 Comparaison des méthodes

Différentes méthodes d'échantillonnage visant à prendre en compte l'information spatiale ont été présentées. Cette partie s'attache à comparer leur efficacité relative, en s'appuyant sur des données réelles.

10.4.1 Le principe

Les données fiscales exhaustives de 2015 sont géoréférencées pour l'ensemble des ménages, ce qui autorise le partitionnement du territoire de la région Provence-Alpes-Côte d'Azur (PACA) en 1 012 unités primaires (UP) d'environ 2 000 résidences principales. Chacune de ces UP est caractérisée par une quinzaine de variables d'intérêt décrivant sa situation socio-économique ou démographique. On s'intéresse aux propriétés statistiques de l'échantillonnage au premier degré de tirage, c'est-à-dire un tirage de m unités primaires parmi les $M = 1\,012$ UP.

On se place dans la situation où l'on ne dispose que des coordonnées géographiques du barycentre des UP au moment de la sélection de l'échantillon. On teste deux jeux de probabilités d'inclusion : le premier correspond à des probabilités constantes, le second à des probabilités proportionnelles au nombre de chômeurs. On teste ces deux jeux pour trois tailles différentes d'échantillons : $m = 30, 60, 100$.

On cherche à évaluer les méthodes présentées précédemment en comparant leurs performances à celles d'une méthode *benchmark* : il s'agit du sondage aléatoire simple pour les plans à probabilités constantes et du tirage systématique trié aléatoirement pour ceux à probabilités inégales. La performance d'une méthode est mesurée à l'aune de deux types d'indicateurs :

1. **variances d'estimateurs de totaux**
   Pour chaque méthode, on cherche à voir dans quelle mesure la variance du total d'une variable

donnée est diminuée par rapport à la variance obtenue avec la méthode *benchmark*. On étudie cela pour un ensemble de variables d'intérêt présentant différents niveaux d'autocorrélation spatiale.
Pour toutes les méthodes sauf celle des plans déterminantaux, les variances des totaux sont estimées par méthode de Monte Carlo, en répliquant 10 000 fois chaque méthode pour chaque jeu de probabilités d'inclusion et chaque taille d'échantillon. Pour les plans déterminantaux, les probabilités d'inclusion double étant connues, la variance peut être calculée de manière exacte.
Puisque l'on souhaite savoir si le gain en variance est plus important quand la variable est autocorrélée spatialement, les 15 variables d'intérêt sont hiérarchisées en fonction de leur niveau d'autocorrélation spatiale, mesuré par l'indice de Moran dilaté par les probabilités d'inclusion. C'est en effet la quantité $\frac{y_i}{\pi_i}$ qui conditionne la qualité des résultats. Lorsque le plan est à probabilités constantes, cela revient à calculer l'indice de Moran directement pour chaque variable (table 10.5).

2. **indicateur dit de Voronoï**
Pour chaque méthode, on calcule également un indicateur empirique de dispersion (dit indice de Voronoï), en suivant STEVENS JR et al. 2004. Son principe est le suivant :
   — on construit le diagramme de Voronoï associé aux seules m UP sélectionnées ;
   — pour une UP i échantillonnée, on identifie, parmi les 1 012 UP d'origine, celles situées dans la cellule associée à i ;
   — on calcule la somme δ_i des probabilités d'inclusion de ces UP. La moyenne des δ_i est égale à 1, puisque la somme des probabilités d'inclusion sur les 1 012 UP vaut m et que les m cellules partitionnent l'espace. Si la procédure a sélectionné peu d'unités autour d'une UP i donnée, δ_i sera supérieur à 1. Si la procédure a sélectionné d'autres unités proches de i, δ_i sera inférieur à 1 (voir figure 10.6) ;
   — pour un échantillon aléatoire $\mathbb{S}$, on définit alors l'indicateur de Voronoï par :

$$\Delta_{\mathbb{S}} = \frac{1}{m-1} \sum_{i \in \mathbb{S}} (\delta_i - 1)^2.$$

Plus une procédure répartit uniformément au plan spatial, plus la dispersion des δ_i mesurée par $\Delta_{\mathbb{S}}$ sera faible. L'espérance de $\Delta_{\mathbb{S}}$ sera estimée par simulation (moyenne sur 10 000 réplications, notée V).

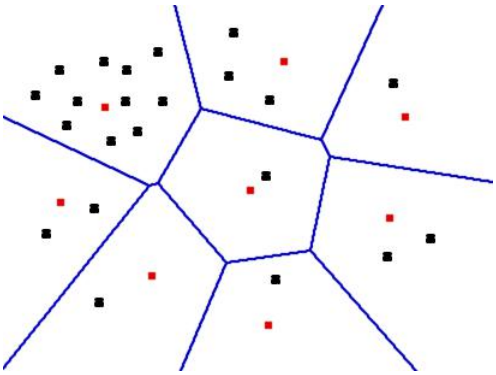


FIGURE 10.6 – Calcul de l'indice de Voronoï
Note : les cellules sont construites autour des unités sélectionnées (rouge). Les δ_i sont calculés sur l'ensemble des unités (rouge et noir)

L'indicateur de Voronoï peut être calculé en R en utilisant la fonction `sb()` du package

BalancedSampling ou à partir des codes R donnés dans BENEDETTI et al. 2015 (pp. 161-162).

Variable	**I de Moran** π constant	**I de Moran dilaté** ($\frac{y_i}{\pi_i}$)
Nombre de ménages percevant des revenus agricoles	0,68	0,66
Revenus salariaux totaux	0,62	0,55
Nombre de couples avec enfant(s)	0,61	0,54
Nombre de bénéficiaires de minima sociaux	0,60	0,61
Nombre de pauvres	0,58	0,58
Nombre d'enfants	0,55	0,52
Nombre d'individus vivant dans un quartier politique de la ville	0,55	0,54
Nombre de ménages propriétaires	0,52	0,47
Niveau de vie total	0,46	0,46
Nombre de chômeurs	0,45	0,42
Nombre de famille monoparentales	0,41	0,43
Nombre d'individus	0,40	0,34
Nombre d'hommes	0,39	0,34
Nombre de femmes	0,24	0,34
Nombre de ménages	0,08	0,38

TABLE 10.5 – Indices de Moran pour différentes variables calculées au niveau des UP de la région PACA
Source : *Insee, Fideli 2015*

10.4.2 Résultats

Une dizaine de méthodes d'échantillonnage spatial sont étudiées :

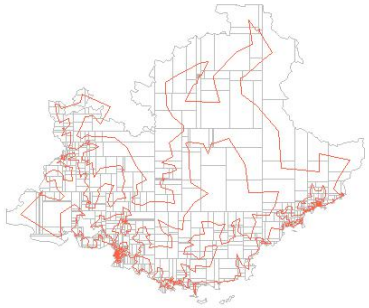
— 4 méthodes de la famille des méthodes de mises à jour itératives des probabilités d'inclusion, dite famille A dans la suite : tirage poissonien, pivot spatial, cube spatial [7], et cube équilibré sur les moments spatiaux ;

— 6 méthodes de la seconde famille, dite famille B dans la suite, fondée sur des tris préalables du fichier. En effet, 3 chemins sont envisagés (figure 10.7) : le chemin du voyageur de commerce (10.7a), celui de Hamilton (10.7b), et GRTS (10.7c), et chacun est suivi d'un tirage systématique ou de la méthode des plans déterminantaux. Les trois chemins sont obtenus à partir d'une méthode exacte et les réplications de tirages d'échantillons ont donc lieu sur un fichier trié de façon unique.

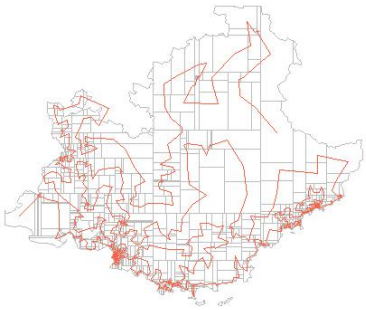
La figure 10.8 représente la quantité $(V^q - V^{ref})/V^{ref}$, où V^q est l'indicateur de Voronoï pour la méthode q et V^{ref} le même indicateur pour le *benchmark*. Une valeur fortement négative révèle une meilleure dispersion spatiale. La figure montre que, pour toutes les méthodes et toutes les tailles d'échantillon, l'indicateur de Voronoï est sensiblement amélioré : de -60 à -70 % par rapport au *benchmark*. La méthode des moments équilibrés est moins performante, tout en restant meilleure que le *benchmark* cependant.

Pour une méthode et une taille d'échantillon données, les figures 10.10 et 10.9 représentent, comme pour l'indicateur de Voronoï, la diminution, par rapport au *benchmark*, de la variance d'une

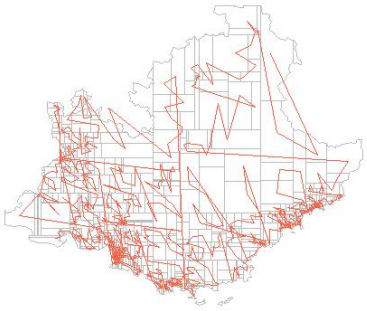
7. Le pivot spatial et le cube spatial sont deux méthodes équivalentes dans ce contexte.



(a) Voyageur de commerce (TSP)



(b) Hamilton



(c) GRTS

FIGURE 10.7 – Chemins reliant les centroïdes des UP
Source : *Insee, Fideli 2015*

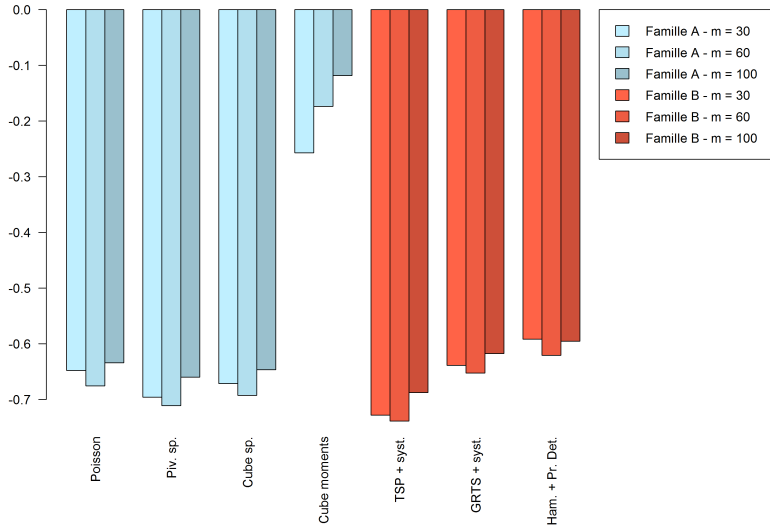


FIGURE 10.8 – $(V^q - V^{ref})/V^{ref}$, où V^q est l'indicateur de Voronoï pour la méthode q et V^{ref} pour le *benchmark*, pour différentes valeurs de m (exemple des probabilités égales)
Source : *Insee, Fideli 2015*
Note : Pour un tirage de 30 UP selon la méthode du tirage poissonien à probabilités égales, l'indicateur de Voronoï (moyenné sur 10 000 réplications) est diminué de 65 % par rapport au sondage aléatoire simple (*benchmark*).

variable d'intérêt. Cette diminution est mise en relation avec l'intensité de l'autocorrélation spatiale de la variable déflatée des probabilités d'inclusion.

Pour les méthodes représentées en figure 10.10, soit la plupart des méthodes étudiées, le gain en termes de variance est d'autant plus fort que la variable est autocorrélée spatialement. Ce résultat est néanmoins plus net à probabilités constantes (10.10a) qu'à probabilités inégales (10.10b). Ces méthodes sont équivalentes en termes de gain. Ainsi le tirage poissonien, le pivot spatial, le cube spatial, et les plans déterminantaux sur fichier trié (tri TSP ou Hamilton), réduisent tous la variance de l'échantillon presque de moitié pour les variables les plus autocorrélées et pour $m = 100$. Par ailleurs, pour toutes les méthodes représentées dans la figure 10.10, le gain relatif en variance est d'autant plus fort que le taux de sondage est élevé. La figure 10.11 illustre ce résultat pour les plans déterminantaux à probabilités constantes.

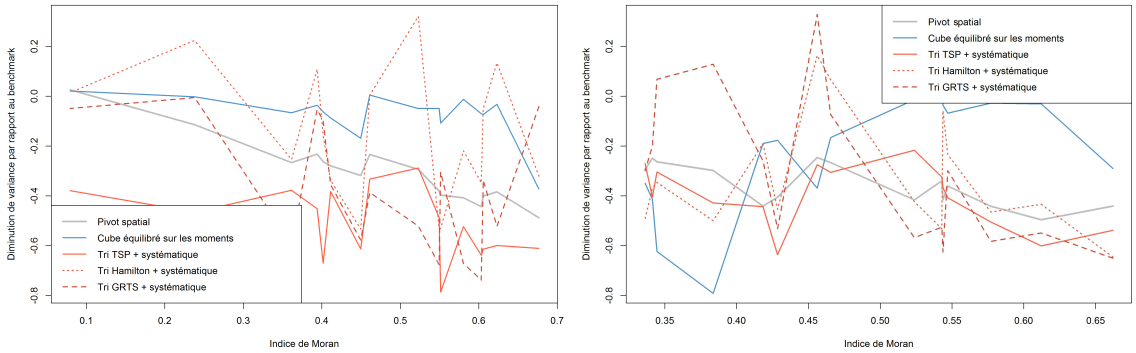
Les quatre méthodes représentées en rouge et bleu dans la figure 10.9 se distinguent par leurs résultats :

— la méthode du cube équilibré sur les moments spatiaux d'ordre 1 et 2 (x, y, $x*y$, x^2 et y^2, où x et y sont les coordonnées spatiales) est, en général, moins performante en termes de gain de variance. En cherchant à se calibrer sur l'inertie de la population totale, cette méthode revient finalement à reproduire dans l'échantillon les regroupements et les éloignements d'unités, de façon antagoniste à la volonté de dispersion de l'échantillon ;
— les tris de fichiers (TSP, Hamilton ou GRTS) suivis d'un tirage systématique, donnent des résultats plus erratiques que les autres méthodes de la même famille. L'entropie [8] du plan de sondage systématique est très faible, et l'est d'autant plus sur un fichier trié de façon unique. Le nombre d'échantillons possibles avec cette méthode est M/m, ce qui explique que ces courbes aient une allure moins lisse et que les conclusions soient plus difficiles à tirer. Ces méthodes restent néanmoins très performantes en termes de dispersion d'échantillon. En particulier, le tri TSP suivi d'un tirage systématique est la méthode qui réduit le plus l'indicateur de Voronoï (figure 10.8). C'est aussi celle qui diminue le plus la variance des variables les plus autocorrélées spatialement. Le tri GRTS, lui, est moins performant, en lien avec une moindre qualité du tri (la longueur totale du chemin obtenu avec GRTS est quasiment deux fois plus grande que le chemin TSP ou Hamilton, voir figure (10.7)).

Conclusion

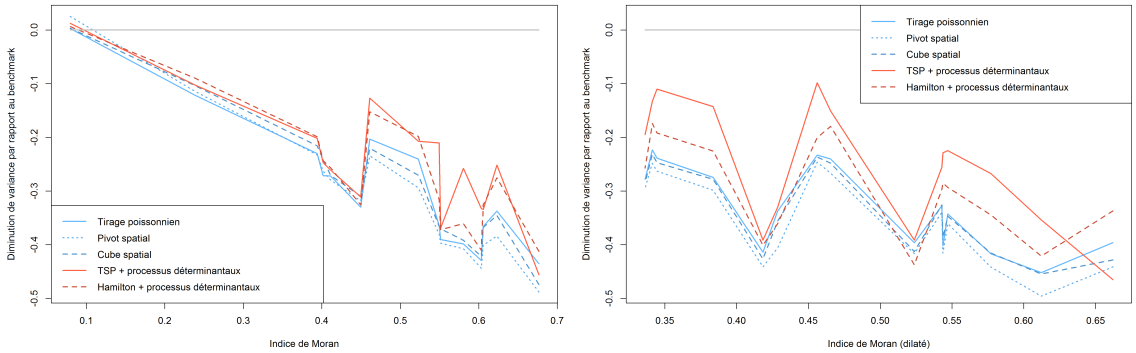
La constitution d'échantillons à partir d'une base de sondage géoréférencée est un nouveau contexte possible de mobilisation judicieuse de l'information géographique. Ce chapitre a présenté différentes méthodes qui utilisent cette information à différents stades de la conception du plan de sondage. En s'appuyant sur des données réelles, nous avons comparé ces différentes méthodes à l'aune d'indicateurs de précision classiques ou originaux, en testant différents jeux de paramètres. La grande majorité des méthodes suggérées s'avèrent efficaces en termes de précision des estimations, même si les méthodes de tirage systématique sur fichier trié apparaissent moins performantes. L'efficacité statistique d'une méthode d'échantillonnage spatial augmente avec le niveau d'autocorrélation spatiale de la variable d'intérêt à estimer.

8. L'entropie est une mesure de désordre. Un plan à forte entropie autorise la sélection d'un grand nombre d'échantillons et laisse donc une place importante à l'aléatoire



(a) Probabilités constantes (b) Probabilités inégales

FIGURE 10.9 – Réductions de variances par rapport au *benchmark* pour différentes méthodes, selon l'indice d'autocorrélation spatiale de la variable (exemple avec $m = 60$)
Note : La méthode du pivot spatial de la figure 10.10 est représentée en trait gris pour comparaison.



(a) Probabilités constantes (b) Probabilités inégales

FIGURE 10.10 – Réductions de variances par rapport au *benchmark* pour différentes méthodes, selon l'indice d'autocorrélation spatiale de la variable (exemple avec $m = 60$)
Source : *Insee, Fideli 2015*
Note : Chaque courbe correspond à une méthode d'échantillonnage spatial, et chaque point de la courbe correspond à 10 000 échantillons tirés selon une même méthode. On représente la variation de la variance d'une variable par rapport à un *benchmark* (en pourcentage), en fonction du niveau d'autocorrélation spatiale de cette variable. Par exemple, pour un tirage à probabilités égales de 60 UP avec la méthode du tirage poissonnien, la variance de la variable "nombre de femmes par UP" (d'indice de Moran 0.24) est diminuée de 11 % par rapport au tirage aléatoire simple

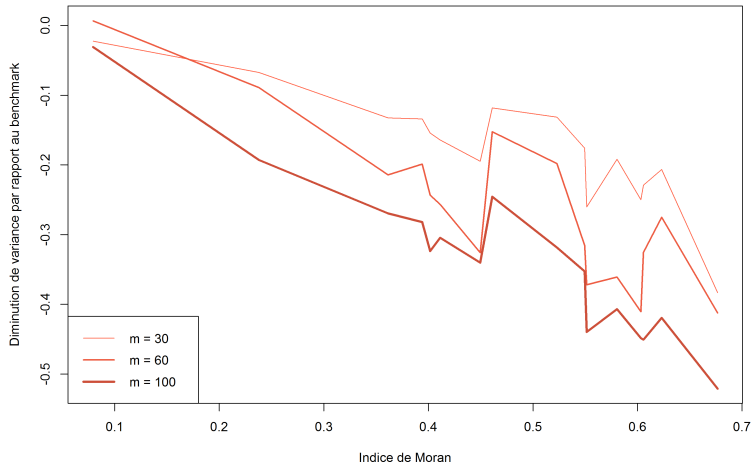


FIGURE 10.11 – Réductions de variances par rapport au *benchmark* selon l'indice d'autocorrélation spatiale de la variable, pour différentes valeurs de m (exemple des processus déterminantaux sur tri de Hamilton, à probabilités constantes)
Source : *Insee, Fideli 2015*
Note : Pour la méthode des processus déterminantaux à probabilités égales, la variance de la variable "nombre de femmes par UP" (d'indice de Moran 0.24) est diminuée de 6,7 % pour un échantillon de 30 UP, de 8,9 % pour un échantillon de 60 UP, et de 19,3 % pour un échantillon de 100 UP, par rapport au tirage aléatoire simple

Références - Chapitre 10

ARDILLY, Pascal (2006). *Les techniques de sondage*. Editions Technip.

BENEDETTI, Roberto, Federica PIERSIMONI et Paolo POSTIGLIONE (2015). *Sampling Spatial Units for Agricultural Surveys*. Springer.

BONDESSON, Lennart et Daniel THORBURN (2008). « A List Sequential Sampling Method Suitable for Real-Time Sampling ». *Scandinavian Journal of Statistics* 35.3, p. 466–483.

BREWER, K.R.W. (1963). « A model of systematic sampling with unequal probabilities ». *Australian & New Zealand Journal of Statistics* 5.1, p. 5–13.

CHAUVET, Guillaume et Ronan LE GLEUT (2017). « Asymptotic results for pivotal sampling with application to spatial sampling ». *Work in progress*.

CONNOR, W.S. (1966). « An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement ». *Journal of the American Statistical Association* 61.314, p. 384–390.

DEVILLE, Jean-Claude et Yves TILLÉ (1998). « Unequal probability sampling without replacement through a splitting method ». *Biometrika* 85.1, p. 89–101.

— (2004). « Efficient balanced sampling : the cube method ». *Biometrika* 91.4, p. 893–912.

DICKSON, Maria Michela et Yves TILLÉ (2016). « Ordered spatial sampling by means of the traveling salesman problem ». *Computational Statistics*, p. 1–14. DOI : 10.1007/s00180-015-0635-1. URL : http://dx.doi.org/10.1007/s00180-015-0635-1.

FAVRE-MARTINOZ, Cyril et Thomas MERLY-ALPA (2017). « Constitution et Tirage d'Unités Primaires pour des sondages en mobilisant de l'information spatiale ». *49èmes Journées de statistique de la Société Française de Statistique*.

GANGANATH, Nuwan, Chi-Tsun CHENG et K Tse CHI (2014). « Data clustering with cluster size constraints using a modified k-means algorithm ». *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on*. IEEE, p. 158–161.

GRAFSTRÖM, Anton (2012). « Spatially correlated Poisson sampling ». *Journal of Statistical Planning and Inference* 142.1, p. 139–147.

GRAFSTRÖM, Anton et J LISIC (2016). « BalancedSampling : Balanced and spatially balanced sampling ». *R package version* 1.2.

GRAFSTRÖM, Anton, Niklas LP LUNDSTRÖM et Lina SCHELIN (2012). « Spatially balanced sampling through the pivotal method ». *Biometrics* 68.2, p. 514–520.

GRAFSTRÖM, Anton et Yves TILLÉ (2013). « Doubly balanced spatial sampling with spreading and restitution of auxiliary totals ». *Environmetrics* 24.2, p. 120–131.

HIDIROGLOU, M.A. et G.B. GRAY (1980). « Algorithm AS 146 : Construction of Joint Probability of Selection for Systematic PPS Sampling ». *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29.1, p. 107–112.

HOUGH, J Ben et al. (2006). « Determinantal processes and independence ». *Probab. Surv* 3, p. 206–229.

LISIC, Jonathan (2015). « Parcel level agricultural land cover prediction ». Thèse de doct. George Mason University.

LISTER, Andrew J et Charles T SCOTT (2009). « Use of space-filling curves to select sample locations in natural resource monitoring studies ». *Environmental monitoring and assessment* 149.1, p. 71–80.

LOONIS, Vincent (2009). « La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation. » *JMS* 2009, p. 23.

LOONIS, Vincent et Xavier MARY (2018). « Determinantal sampling designs ». *Journal of Statistical Planning and Inference*.

MADOW, William G (1949). « On the theory of systematic sampling, II ». *The Annals of Mathematical Statistics*, p. 333–354.

MALINEN, Mikko I et Pasi FRÄNTI (2014). « Balanced k-means for clustering ». *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, p. 32–41.

MARKER, David A et Don L STEVENS (2009). « Sampling and inference in environmental surveys ». *Handbook of Statistics* 29, p. 487–512.

PINCIARO, Susan J (1978). « An algorithm for calculating joint inclusion probabilities under PPS systematic sampling ». *of : ASA Proceedings of the Section on Survey Research Methods*, p. 740–740.

STEVENS JR, Don L et Anthony R OLSEN (2004). « Spatially balanced sampling of natural resources ». *Journal of the American Statistical Association* 99.465, p. 262–278.

TAI, Chen-Ling et Chen-Shu WANG (2017). « Balanced k-Means ». *Asian Conference on Intelligent Information and Database Systems*. Springer, p. 75–82.