Échantillonnage doublement équilibré avec étalement spatial et restitution de variables auxiliaires

Anton Grafström et Yves Tillé

Swedish University of Agricultural Sciences

Université de Neuchâtel

Colloque francophone sur les sondages

Bruz 2012

Échantillon

1. Un échantillon est un vecteur aléatoire.
2. Chaque composante de ce vecteur est le nombre de fois que l'unité est sélectionnée

Exemple

Population $U = \{1, 2, 3, 4, 5\}$

Échantillon comme un sous-ensemble $s = \{2, 4\}$

Échantillon comme un vecteur $\mathbf{I} = (I_1, I_2, I_3, I_4, I_5) = (0, 1, 0, 1, 0)$.

Plan de sondage

Définition

A plan de sondage est une distribution de probabilité sur les sous-ensembles de la population

$$p(s) \geq 0, \sum_{s \subset S} p(s) = 0.$$

Exemples : plans simples, stratifiés, à probabilités inégales, équilibrés.

S est l'échantillon aléatoire $\Pr(S = s) = p(s)$.

L'échantillon aléatoire peut aussi être vu comme un vecteur aléatoire positif discret

Définition

Probabilités d'inclusion $\pi_k = \Pr(k \in S) = \mathrm{E}(I_k), k \in U$.

Probabilités d'inclusion d'ordre deux $\pi_{k\ell} = \Pr(k \text{ et } \ell \in S) = \mathrm{E}(I_k I_\ell)$.

Plan de sondage

Total de la variable d'intérêt $Y = \sum_{k \in U} y_k$. (y_k pas aléatoire)

Définition

Estimateur de Horvitz-Thompson (HT)

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k.$$

L'estimateur HT est sans biais ssi $\pi_k > 0$ pour tout $k \in U$.

$$var(\widehat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell).$$

(avec $\pi_{kk} = \pi_k$).

Information auxiliaire

- Vecteur d'informations auxiliaires $\mathbf{x}_k \in \mathbb{R}^p$ connu sur toutes les unités de la population (registre).
- Utilisation d'informations auxiliaires
  - À l'étape de la collecte : échantillonnage équilibré.
  - À l'étape de l'estimation : calage.
- L'échantillonnage équilibré consiste à sélectionner un échantillon aléatoire avec des probabilités d'inclusion données π_k tel que

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} \approx \sum_{k \in U} \mathbf{x}_k.$$

  Les estimateurs HT des totaux des variables auxiliaires sont égaux ou presqu'égaux aux vrais totaux de la population.
- Cas particuliers : plans de taille fixe, stratification.

Échantillonnage équilibré

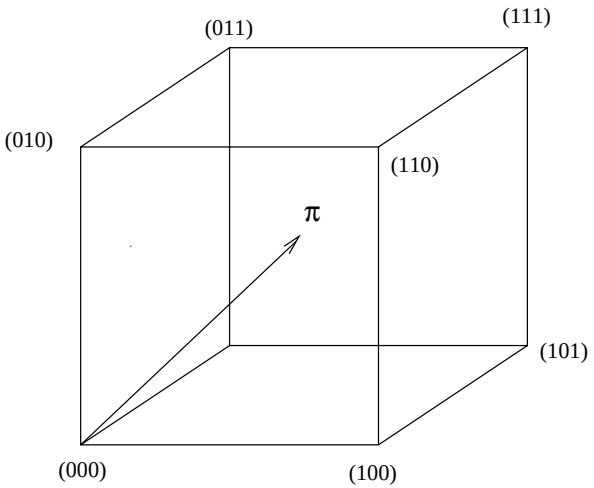
- L'algorithme du cube (Deville & Tillé, 2004; Tillé, 2006) permet de sélectionner des échantillons équilibrés dans des registres. (Les estimateurs des totaux des variables du registre sont égaux aux totaux de la population).
- Modèle linéaire $y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k$ avec $\mathrm{var}(\varepsilon_k) = \sigma_k^2$ et $\mathrm{cov}(\varepsilon_k, \varepsilon_\ell) = 0$.
- But : Minimiser l'Erreur Quadratique Moyenne Anticipée (MSE)

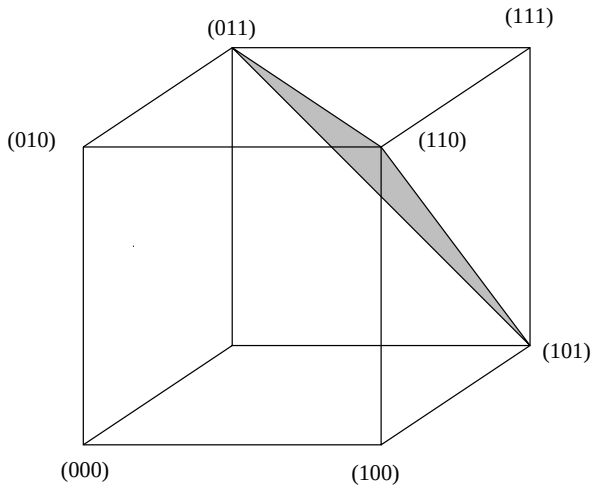
$$\mathrm{MSE}(\widehat{Y}) = \mathrm{E}_p \mathrm{E}_M (\widehat{Y} - Y)^2.$$

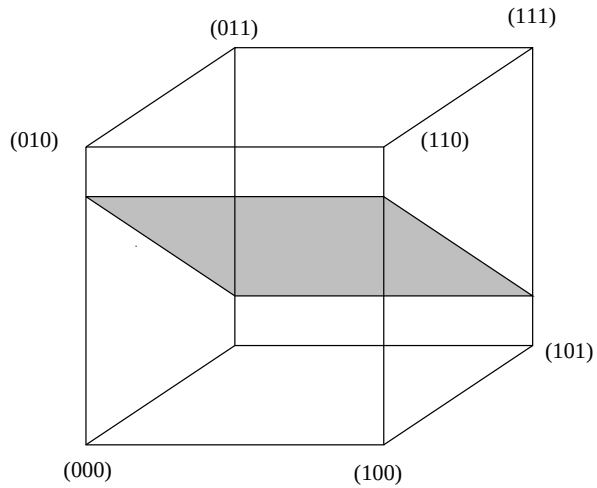
- Nedyalkova & Tillé (2008) : la stratégie optimale (plan de sondage + estimateur) consiste à
  - équilibrer l'échantillon sur $\mathbf{x}_k$,
  - utiliser des probabilités d'inclusion $\pi_k \propto \sigma_k$,
  - utiliser l'estimateur HT.

Principe de la méthode du cube

- Chaque échantillon est un sommet d'un hypercube dans $\mathbb{R}^N$.
- Le vecteur des probabilités d'inclusion est à l'intérieur de cet hypercube.
- Les contraintes d'équilibrage définissent un sous-espace.







Idée de l'algorithme, phase de vol

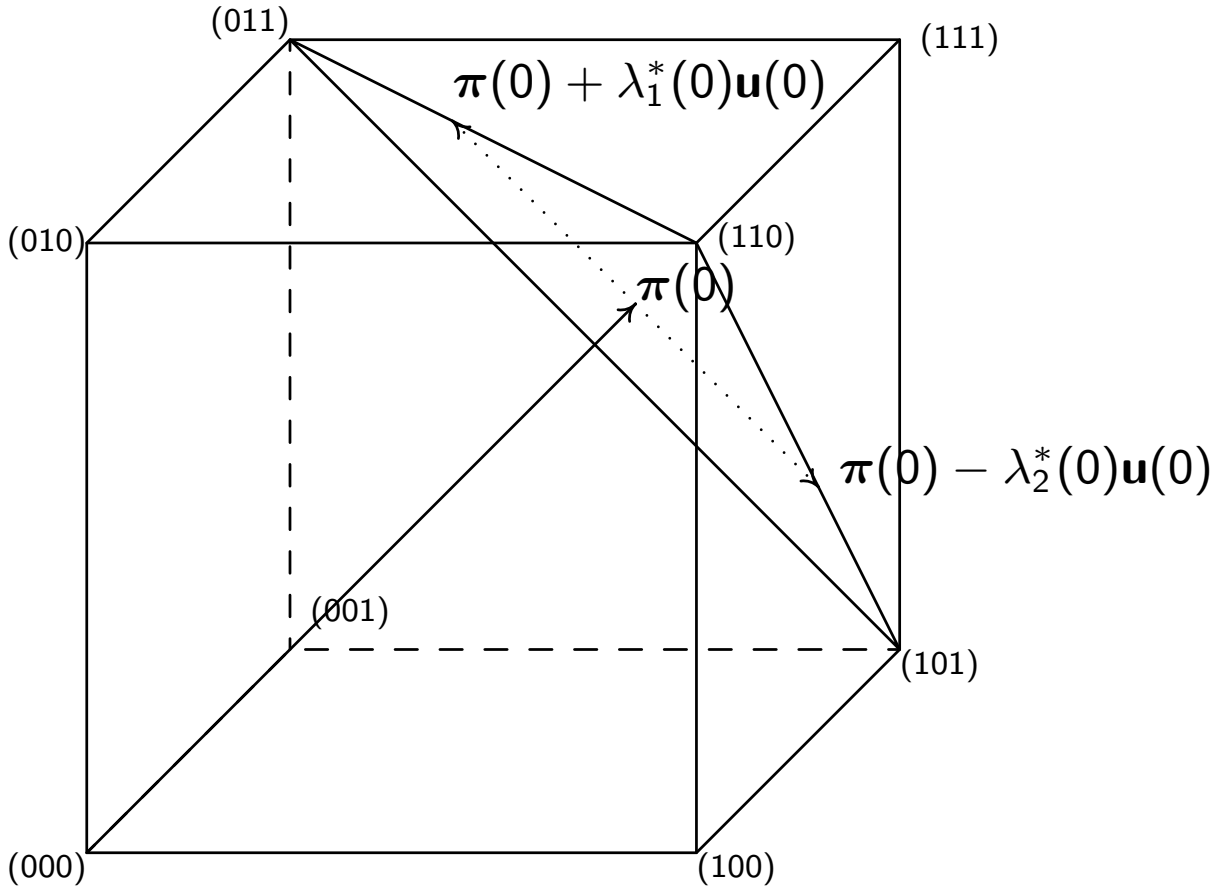


FIGURE: Phase de vol dans une population de taille $N = 3$ avec une contrainte de taille fixe $n = 2$

Principe de la méthode du cube

- Deux phases : phase de vol et phase d'atterrissage.
- Phase de vol : promenade aléatoire dans l'intersection du cube et du sous-espace des contraintes jusqu'à un sommet.
- Si ce sommet n'est pas un échantillon, phase d'atterrissage. (les contraintes sont relâchées).

Échantillonnage spatial

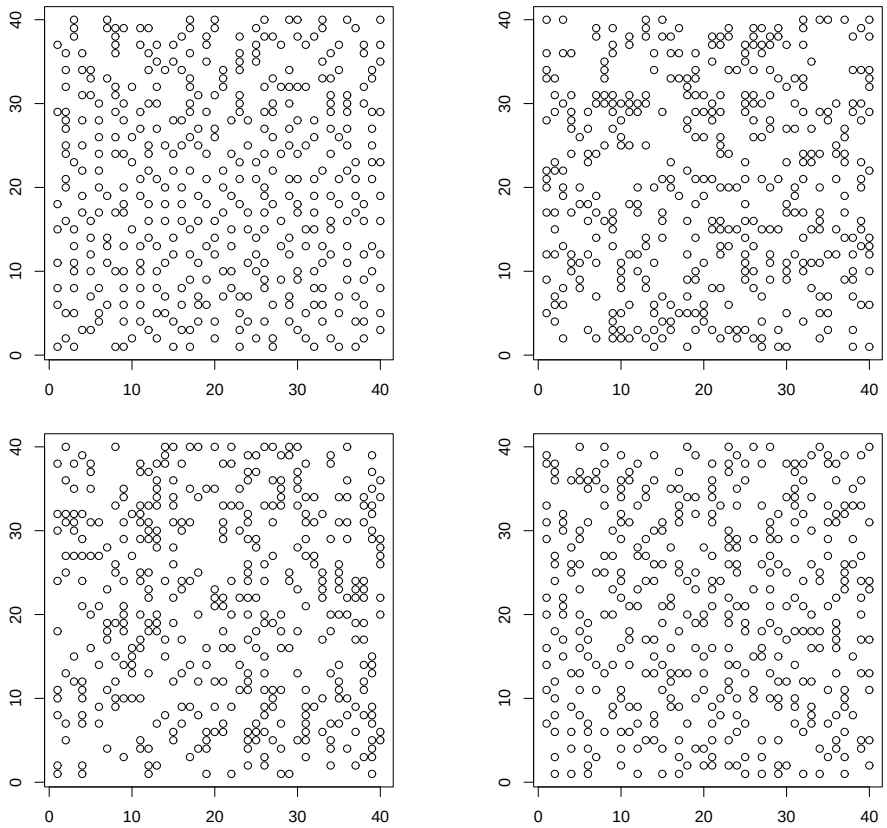
- Avec des données spatiales, on a généralement de l'autocorrélation

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k, \text{ pour tous } k \in U,$$

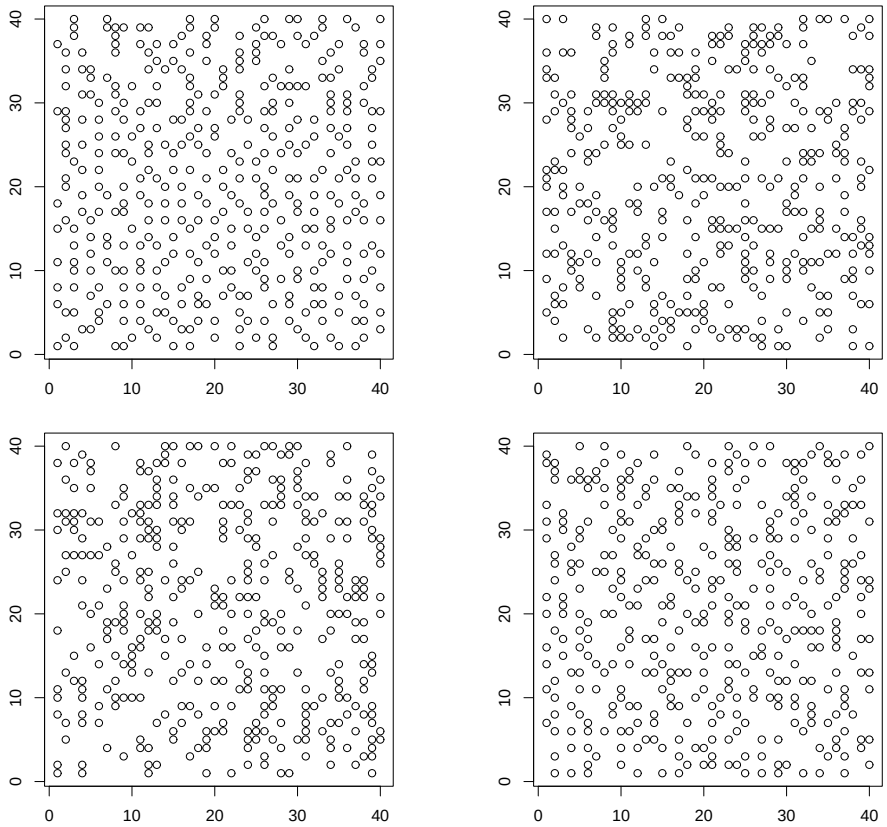
avec $\mathrm{var}_M(\varepsilon_k) = \sigma_k^2$, $\mathrm{cov}_M(\varepsilon_k, \varepsilon_\ell) = \sigma_k \sigma_\ell \rho_{k\ell}$,

- On suppose que $\rho_{k\ell}$ devient plus grand quand k et ℓ sont proches.
- Un plan de sondage très efficace (par rapport à l'erreur quadratique moyenne anticipée MSE) consiste à :
  - utiliser un échantillonnage équilibré sur les variables $\mathbf{x}_k$,
  - éviter la sélection d'unités voisines, i.e. sélectionnée un échantillon bien étalé (ou équilibré spatialement),
  - utiliser des probabilités d'inclusion proportionnelles à σ_k.
  - utiliser l'estimateur HT.

Échantillonnage spatial



Échantillonnage spatial



(a) Plan 1 (étalé et équilibré), (b) Plan 2 (seulement équilibré),
(c) Plan 3 (simple), (d) Plan 4 (seulement étalé)

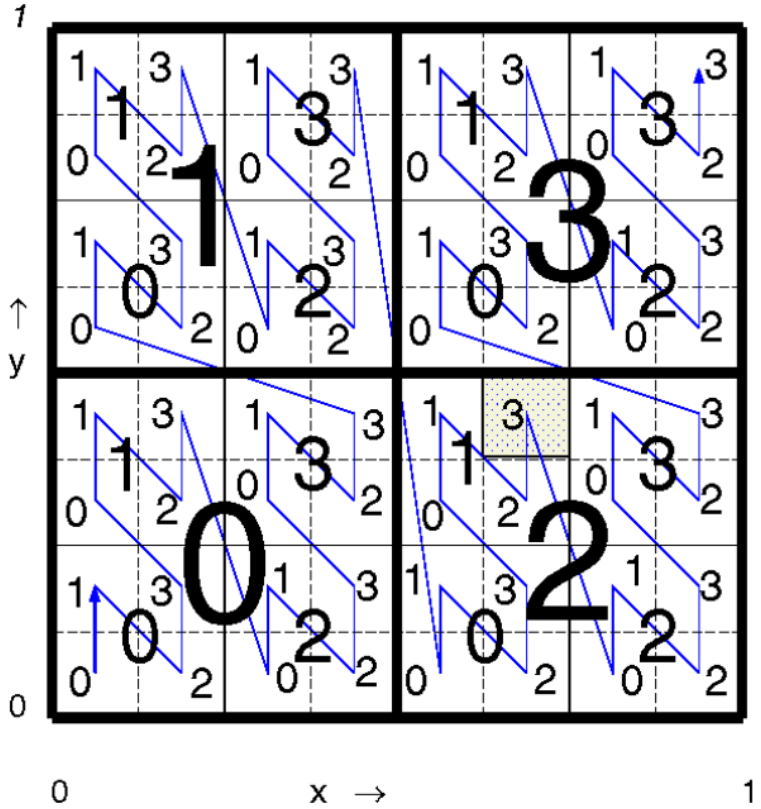
Generalized Random Tessellation Sampling

Algorithme de Stevens & Olsen (2003, 2004)

1. On crée une grille hiérarchisée avec des adresses.
2. On randomise les adresses.
3. On trie les unités sur une ligne selon les adresses.
4. On réalise un tirage systématique le long de cette ligne.

L'échantillon est bien étalé mais les totaux ne sont pas équilibrés.

Generalized Random Tessellation Sampling



L'échantillon est bien étalé mais les totaux ne sont pas équilibrés.

La méthode du pivot Deville & Tillé (1998)

Méthode pour sélectionner un échantillon avec des probabilités d'inclusion inégales.

Soient i et j deux unités telles que $0 < \pi_i, \pi_j < 1$.

Seulement ces deux unités sont modifiées à chaque étape (π_i, π_j) avec la règle suivante. Si $\pi_i + \pi_j < 1$, alors

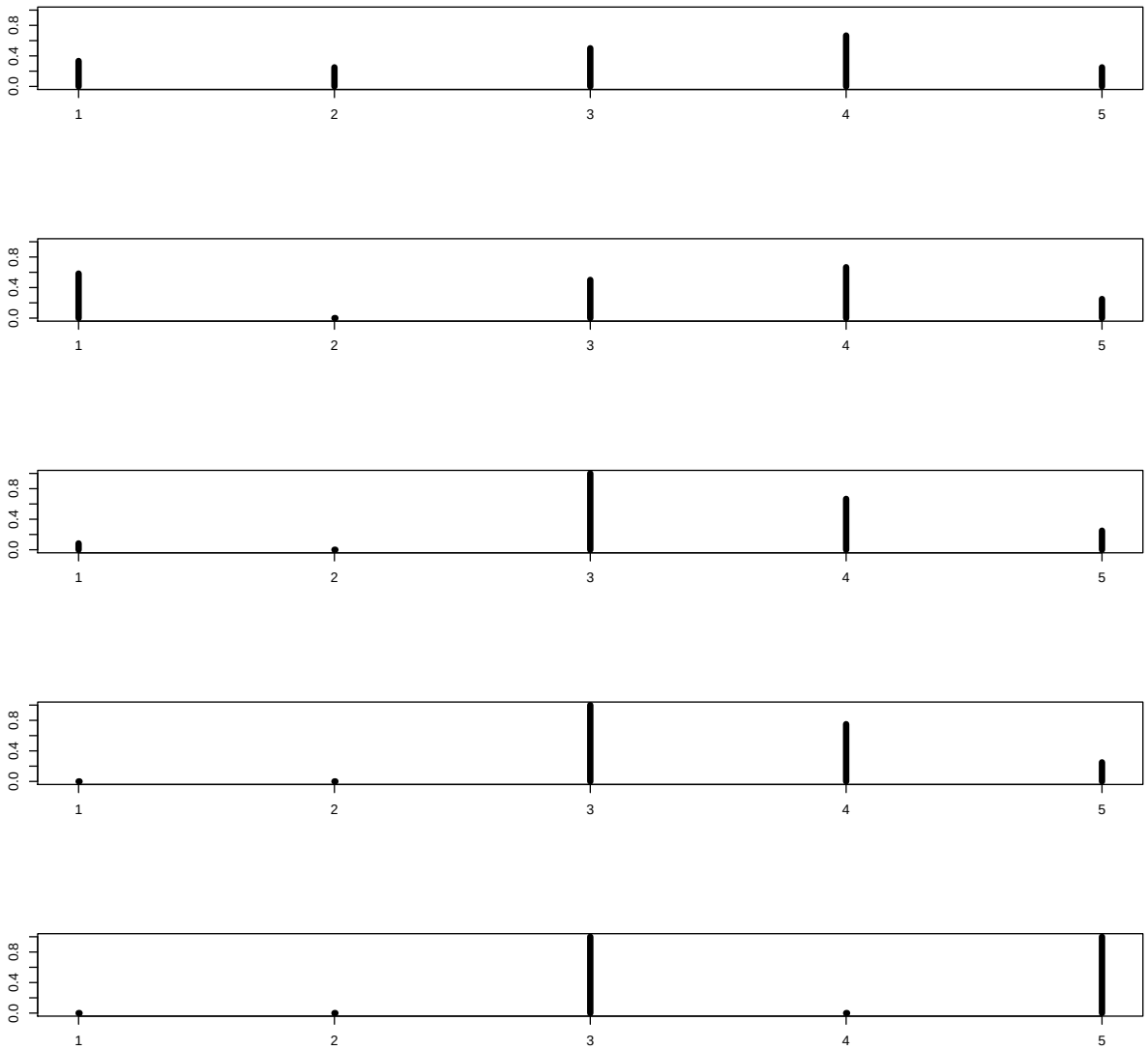
$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) & \text{avec la probabilité } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) & \text{avec la probabilité } \frac{\pi_i}{\pi_i + \pi_j} \end{cases},$$

et si $\pi_i + \pi_j \geq 1$, alors

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

Cette simple procédure est répétée jusque toutes les valeurs soient soit égales à 1 soit à 0.

Exemple méthode du pivot



Méthode du pivot local

Algorithme de Grafström et al. (2012)

1. On choisit deux unités i et j avec des probabilités strictement entre 0 et 1 qui sont proches spatialement.
2. On applique une étape de la méthode du pivot uniquement sur i et j.
3. On répète ces deux étapes.

L'échantillon est bien étalé mais n'est pas équilibré.

Algorithme pour un échantillonnage à la fois étalé et équilibré (doublement équilibré)

- Soit p le nombre de variables d'équilibrage.
- Avec la méthode du cube, la dimension de l'espace des contraintes est égale à $N - p$.
- Pour exécuter une étape de la phase de vol de la méthode du cube, la taille de la population doit être au moins égale à $p + 1$.

Algorithme On répète ces étapes :

(1) On sélectionne un ensemble de $p + 1$ unités voisines qui ont des probabilités strictement entre 0 et 1.

(2) On exécute une étape de la phase de vol.

On réalise la phase d'atterrissage.

Exemple avec le jeu de données 'Meuse'

Pebesma (2004) : 'Ces données contiennent les coordonnées et les concentrations en métaux lourds (ppm) dans la plaine alluviale de la Meuse.'

- x, x-coordonnée topographique,
- y, y-coordonnée topographique,
- cadmium, concentration en cadmium,
- copper, concentration en cuivre,
- lead, concentration en plomb,
- zinc, concentration en zinc,
- elev, altitude relative,
- om, pourcentage de matière organique.

Exemple avec le jeu de données 'Meuse'

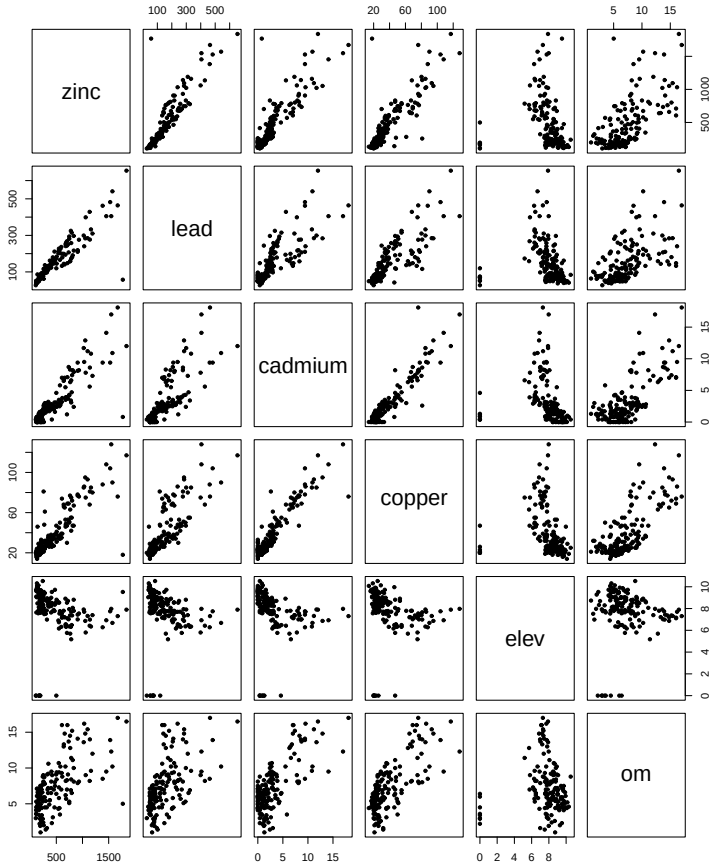


FIGURE: Relations entre les variables d'équilibrage (copper, elev, om) et les variables d'intérêt (zinc, lead, cadmium)

Exemple avec le jeu de données 'Meuse'

On compare 5 plans :

1. Échantillonnage étalé et équilibré. (Nouvelle méthode).
2. Échantillonnage équilibré par la méthode du cube. Dans ce cas, l'échantillon n'est pas étalé et les coordonnées topographiques ne sont pas prises en compte.
3. Échantillonnage à probabilités inégales sans remise. Si les probabilités d'inclusion sont égales, on obtient un plan simple sans remise de taille fixe.
4. Méthode du pivot local. Échantillonnage étalé, mais les variables d'équilibrage ne sont pas utilisées.
5. Generalized Random-Tessellation Stratified (GRTS). Échantillonnage étalé, mais les variables d'équilibrage ne sont pas utilisées.

Exemple avec le jeu de données 'Meuse'

Deux ensembles de probabilités d'inclusion :

- Probabilités égales.
- Probabilités d'inclusion proportionnelles à la concentration en cuivre.

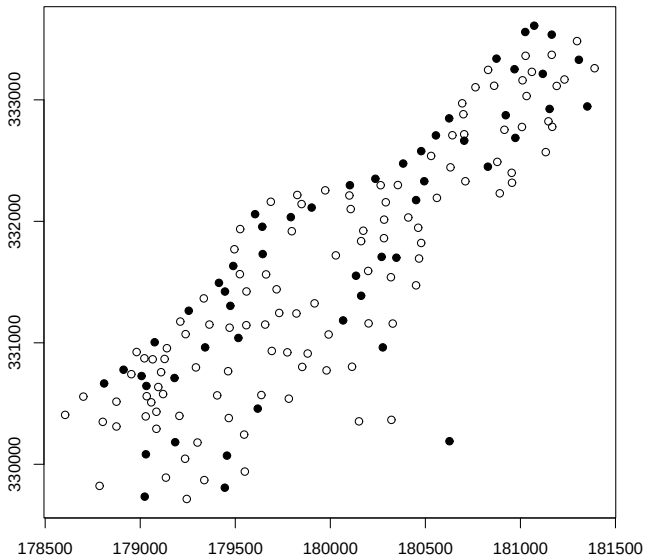


FIGURE: Échantillon de 50 unités sélectionnés dans le jeu de données 'Meuse'. L'échantillon est sélectionné avec des probabilités d'inclusion inégales proportionnelles à la concentration en cuivre. Le plan de sondage est étalé et équilibré.

TABLE: Résultats de 10,000 simulations avec des probabilités d'inclusion égales

Variance approchée par simulations			
	zinc	plomb	cadmium
Étalé et équilibré	12116860	964038	715
Méthode du cube	15386870	1594422	783
Plan simple	51896830	4448526	4438
Méthode du pivot local	33179400	2640523	3001
GRTS	37565209	2932201	3256
Varance approchée par simulations par rapport au plan simple			
	zinc	plomb	cadmium
Étalé et équilibré	23.35%	21.67%	16.11%
Méthode du cube	29.65%	35.84%	17.64%
Plan simple	100.00%	100.00%	100.00%
Méthode du pivot local	63.93%	59.36%	67.62%
GRTS	72.38%	65.91%	73.37%

TABLE: Résultats de 10.000 simulations avec des probabilités d'inclusion inégales

Varance approchée par simulations			
	zinc	plomb	cadmium
Étalé et équilibré	19483080	501241	328
Méthode du cube	22022460	854072	400
Échantillonnage à probabilités inégales	21547960	901392	779
Méthode du pivot local	19915120	571500	601
GRTS	19502579	575623	586

Varance approchée par simulations par rapport au plan simple			
	zinc	plomb	cadmium
Étalé et équilibré	37.54%	11.27%	7.39%
Méthode du cube	42.44%	19.20%	9.01%
Échantillonnage à probabilités inégales	41.52%	20.26%	17.55%
Méthode du pivot local	38.37%	12.85%	13.54%
GRTS	37.58%	12.94%	13.20%

Conclusions

- La méthode de Grafström marche aussi bien que la méthode GRTS.
- La nouvelle méthode est la meilleure, car elle combine étalement et équilibrage.
- Application : sélection de sites pour le monitoring de prairies sèches en Suisse, sélection de sites pour l'enquête sur les libellules.

References

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

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

Grafström, A., Lundström, N. & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics, DOI : 10.1111/j.1541-0420.2011.01699.x* .

Nedyalkova, D. & Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika* **95**, 521–537.

Pebesma, E. J. (2004). Multivariable geostatistics in s : the gstat package. *Computers & Geosciences* **30**, 683–691.

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

Stevens, D. L. J. & Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *EnvironMetrics* **14**, 593–610.

Tillé, Y. (2006). *Sampling Algorithms*. New York : Springer.