Plans de sondage déterminantaux et échantillonnage spatial

Vincent Loonis[1] & Xavier Mary [2]

[1] *Insee, Division des Méthodes et des Référentiels Géographiques, Paris, France : vincent.loonis@insee.fr*

[2] *Université Paris Nanterre, Modal'X, France : xavier.mary@u-paris10.fr*

Résumé. L'échantillonnage spatial dans une population finie a fait l'objet de nombreux travaux au cours des dernières années. La stratégie d'ensemble est répulsive. Elle consiste à attribuer une faible probabilité d'inclusion double à deux unités proches *géographiquement*. La propriété de répulsivité apparaît naturellement dans certains processus ponctuels, dont les processus determinantaux. L'application de ces processus au domaine des sondages met en évidence le rôle des matrices de projection orthogonale contractantes dont la diagonale correspond à un jeu de probabilités fixées a priori. La construction effective de telles matrices, qui fera l'objet de la première partie de la présentation, autorise, dans une seconde partie, la comparaison des performances des plans déterminantaux avec celles des procédures d'échantillonnage spatial déjà existantes.

Mots-clés. Processus déterminantal, échantillonnage spatial ...

Introduction

L'échantillonnage spatial dans une population finie a fait l'objet de nombreux travaux au cours des dernières années (Stevens Jr and Olsen (2004), Grafström (2012),Grafström and Tillé (2013),Benedetti et al. (2015),Dickson and Tillé (2016)). La stratégie d'ensemble est répulsive. Elle consiste à attribuer une faible probabilité d'inclusion double à deux unités proches *géographiquement*. La propriété de répulsivité apparaît naturellement dans certains processus ponctuels, dont les processus determinantaux (Soshnikov (2000), Hough et al. (2006), Kulesza (2012)). L'application de ces processus au domaine des sondages a été étudiée dans Loonis and Mary (2015). Elle met en évidence l'importance de la constrution de matrices de projection orthogonale contractante dont la diagonale correspond à un jeu de probabilités fixées a priori. La construction effective de telles matrices s'appuyant sur les travaux de Schur (1911), Horn (1954), Kadison (2002), Fickus et al. (2013) autorise la comparaison des performances des plans déterminantaux avec les procédures d'échantillonnage spatial déjà existantes.

1 Les plans de sondage déterminantaux

Définition 1.1 (Plan de sondage déterminatal) *Un plan de sondage $\mathcal{P}$ sur une population finie U est determinantal si il existe une matrice hermitienne K indexée par U, appelée noyau, telle que pour tout $s \in 2^U$, $\sum_{s' \supseteq s} \mathcal{P}(s') = \det(K_{|s})$. Un tel plan sera noté $DSD(K)$. Une variable aléatoire $\mathbb{S}$ à valeurs dans 2^U de loi $DSD(K)$ est un échantillon aléatoire déterminantal (de noyau K). Elle vérifie que, pour tout $s \in 2^U$,*

$$pr(s \subseteq \mathbb{S}) = \det(K_{|s}).$$

On notera $\mathbb{S} \sim DSD(K)$

Macchi (1975) and Soshnikov (2000) prouvent qu'une matrice hermitienne K definit un processus ponctuel déterminantal, et donc un $DSD(K)$, si et seulement si$K$ est une matrice contractante, c'est à dire une matrice dont les valeurs propres sont dans l'intervalle $[0, 1]$. Les probabilités d'inclusion simple et double se déduisent de la définition précédente. On les note π_k et π_{kl}. Sous forme matricielle, on a :

$$\begin{aligned}
\pi_k &= pr(k \in \mathbb{S}) = K_{kk}, && (1)\\
\pi_{kl} &= pr(k, l \in \mathbb{S}) = K_{kk}K_{ll} - \mid K_{kl} \mid^2 \ (k \neq l), && (2)\\
\Delta_{kl} &= \begin{cases} \pi_{kl} - \pi_k\pi_l = - \mid K_{kl} \mid^2 \ (k \neq l), \\ \pi_k(1 - \pi_k) = K_{kk}(1 - K_{kk}) \ (k = l). \end{cases} && (3)
\end{aligned}$$

soit

$$\Delta = \overline{(I_N - K)} * K = (I_N - K) * \overline{K}, \quad (4)$$

où $*$ est le produit matriciel d'Hadamard (produit terme à terme), et $|z|$ désigne le module du nombre complexe z.

Proposition 1.1 *D'après* (3) *un plan de sondage déterminantal vérifie les conditions de Sen-Yates-Grundy:*

$$\pi_{kl} \leq \pi_k\pi_l \ (k \neq l). \quad (5)$$

Un résultat particulièrement important dans ce cadre est dû à Hough et al. (2006)

Thèorème 1.1 (Taille de l'échantillon (1)) *Soit $\mathbb{S} \sim DSD(K)$, la variable aléatoire $\sharp\mathbb{S}$, cardinal de $\mathbb{S}$, a même loi que celle de la somme de N variables de Bernoulli indépendantes $B_1, \cdots, B_N$ des paramètres $\lambda_1, \cdots, \lambda_N$, éléments du spectre $Sp(K)$ de K.*

Corollaire 1.1 (Taille de l'échantillon (2)) *Let $\mathbb{S} \sim DSD(K)$. Soit*

1. $E(\sharp\mathbb{S}) = tr(K)$.

2. $var(\sharp\mathbb{S}) = tr(K - K^2) = \sum_{k \in Sp(K)} \lambda_k(1-\lambda_k) = \sum_{k,l \in U} \Delta_{kl}$.

3. $pr(\mathbb{S} = \emptyset) = 0$ *iff* $1 \in Sp(K)$.

4. $DSD(K)$ *est un plan de sondage déterminantal de taille fixe si et seulement si* K *is une matrice de projection.*

2 Construction de plans déterminantaux de probabilités fixées a priori

Les parties précédentes montrent l'importance des matrices de projection contractantes de diagonale fixée dans la mise en place des plans déterminantaux de probabilités d'inclusion données a priori et de taille fixe. L'existence de telles matrices de diagonale et spectre fixés est assurée (Schur (1911),Horn (1954)), mais les preuves sont non constructives. En s'appuyant sur des résultats de Kadison (2002), on montre cependant les résultats ci-après.

Soit Π un vecteur de taille N tel que $0 < \Pi_k < 1$ et $\sum_{k=1}^{k=N} \Pi_k = n \in \mathbb{N}^*$. on pose $k_0 = 0$ et pour tout entier r tel que $1 \leq r \leq n$, on définit (voir figure 1)

- $1 < k_r \leq N$ un entier tel que $\sum_{k=1}^{k_r-1} \Pi_k < r$ et $\sum_{k=1}^{k_r} \Pi_k \geq r\}$;
- $\alpha_{k_r} = r - \sum_{k=1}^{k_r-1} \Pi_k$ et $\alpha_k = \Pi_k$ si $k \neq k_r$;
- $\gamma_r^{r'} = \sqrt{\prod_{j=r+1}^{r'} \frac{(\Pi_{k_j} - \alpha_{k_j})\alpha_{k_j}}{(1-\alpha_{k_j})(1-(\Pi_{k_j}-\alpha_{k_j}))}}$ for $r < r'$, $\gamma_r^{r'} = 1$ autrement.

On construit la matrice symétrique réelle P^{Π} telle que:

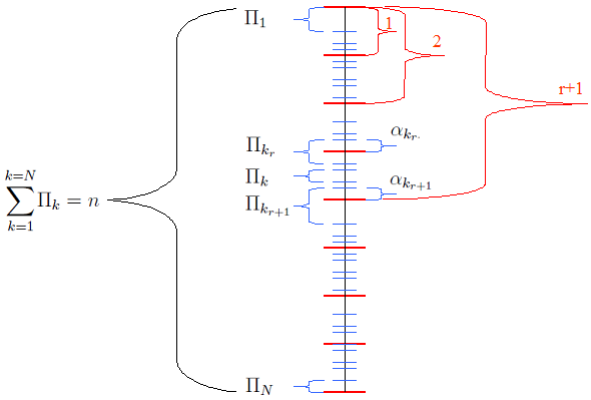
- pour tout $1 \leq k \leq N$, $P^{\Pi}_{kk} = \Pi_k$;
- pour tout $k > l$, P^{Π}_{kl} prend les valeurs définies dans le tableau 1:

Table 1: Valeurs de P_{kl}^{Π} : $k > l$

	Valeurs de l	
Valeurs de k	$l = k_r$	$k_r < l < k_{r+1}$
$k_{r'} < k < k_{r'+1}$	$-\sqrt{\Pi_k}\sqrt{\frac{(1-\Pi_l)(\Pi_l-\alpha_l)}{1-(\Pi_l-\alpha_l)}}\gamma_r^{r'}$	$\sqrt{\Pi_k\Pi_l}\gamma_r^{r'}$
$k = k_{r'+1}$	$-\sqrt{\frac{(1-\Pi_k)\alpha_k}{1-\alpha_k}}\sqrt{\frac{(1-\Pi_l)(\Pi_l-\alpha_l)}{1-(\Pi_l-\alpha_l)}}\gamma_r^{r'}$	$\sqrt{\frac{(1-\Pi_k)\alpha_k}{1-\alpha_k}}\sqrt{\Pi_l}\gamma_r^{r'}$

Thèorème 2.1 (Loonis and Mary (2016))
La matrice P^{Π} est une matrice de projection, et $DSD(P^{\Pi})$ est un plan déterminantal de taille fixe de probabilités d'inclusion $\pi_k = \Pi_k, 1 \leq k \leq N$.

Figure 1: Représentation graphique des quantités intervenant dans le théorème 2.1



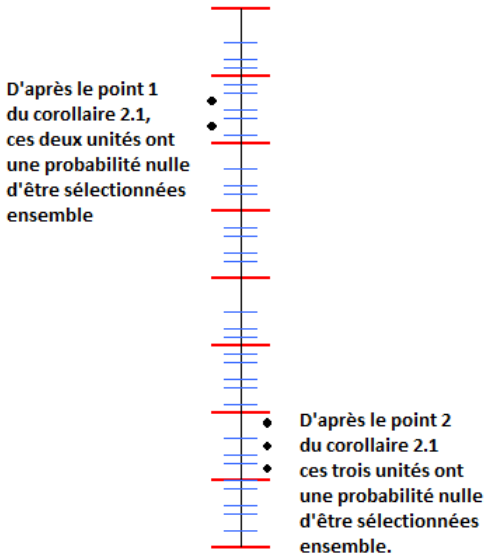
Corollaire 2.1 *Soit P^{Π} la matrice précédente et $DSD(P^{\Pi})$ le plan de sondage associé.*

1. *Si $(k,l) \in]k_r+1, k_{r+1}-1[^2$ alors $\pi_{kl} = 0$.*
2. *Si $i \in]k_r+1, k_{r+1}-1[$, $j = k_{r+1}$, $k \in]k_{r+1}+1, k_{r+2}-1[$ alors $\pi_{ijk} = 0$.*
3. *Soit $B_r = [1; k_r+1]$, alors les valeurs propres de $K_{|B_r}$ sont 1 de multiplicité r et 0 de multiplicité $k_r - r$: l'échantillon aléatoire $\mathbb{S}$ a r ou $r+1$ éléments dans B_r $(r \leq \sharp(\mathbb{S} \cap B_r) \leq r+1)$.*
4. *Si $k-l$ est grand alors $P_{kl}^{\Pi} \approx 0$, les événements $\{k \in \mathbb{S}\}$ et $\{l \in \mathbb{S}\}$ sont asymptotiquement independants. En pratique, $\pi_{kl} \approx \Pi_k\Pi_l$ semble valide même pour de petites valeurs de $k-l$ en valeur absolue.*

5. *Soit* $r_1, \cdots, r_H$ *l'ensemble des valeurs de* $1 \leq r \leq n$ *telles que* $\sum_{k=1}^{k_r} \Pi_k = r$, *et soit* $r_0 = 0$, *alors* $DSD(P^{\Pi})$ *est un plan stratifié ayant* H *strates* $]k_{r_{h-1}}, k_{r_h}]$. *En particulier, si les* Π_k *sont constants et que* n *divise* N *alors le plan consiste à sélectionner exactement une unité dans chaque strate à probabil'e constante.*

Afin d'aider à mieux comprendre comment l'échantillon est sélectionné, la figure 2 illustre graphiquement certaines des propriétés précédentes.

Figure 2: Représentation graphique des propriétés du corollaire 2.1



On exhibe enfin deux exemples de matrices construites selon la méthode précédente.

Exemple 2.1 *Soit* $\Pi = (\frac{1}{2}, \frac{3}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5})^T$ *and* $\Pi' = (\frac{1}{2}, \frac{1}{5}, \frac{3}{4}, \frac{4}{5}, \frac{2}{5}, \frac{3}{5}, \frac{3}{4})^T$. *On notera que* Π' *est une permutation de* Π, *et que* $\Pi_1 + \Pi_2 + \Pi_3 = 2$. *Alors*

$$P^{\Pi} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & 0 & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{3}{4} & -\frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{5} & \frac{\sqrt{2}}{5} & \frac{2}{5\sqrt{3}} & \frac{\sqrt{2}}{5\sqrt{3}} \\ 0 & 0 & 0 & \frac{\sqrt{2}}{5} & \frac{2}{5} & \frac{2\sqrt{2}}{5\sqrt{3}} & \frac{2}{5\sqrt{3}} \\ 0 & 0 & 0 & \frac{2}{5\sqrt{3}} & \frac{2\sqrt{2}}{5\sqrt{3}} & \frac{3}{5} & -\frac{\sqrt{2}}{5} \\ 0 & 0 & 0 & \frac{\sqrt{2}}{5\sqrt{3}} & \frac{2}{5\sqrt{3}} & -\frac{\sqrt{2}}{5} & \frac{4}{5} \end{pmatrix},$$

$$P^{\Pi'} = \begin{pmatrix} \frac{1}{2} & \frac{1}{\sqrt{10}} & \frac{\sqrt{3}}{2\sqrt{14}} & \frac{\sqrt{3}}{\sqrt{70}} & \frac{1}{\sqrt{35}} & \frac{1}{\sqrt{65}} & \frac{1}{2\sqrt{26}} \\ \frac{1}{\sqrt{10}} & \frac{1}{5} & \frac{\sqrt{3}}{2\sqrt{35}} & \frac{\sqrt{3}}{5\sqrt{7}} & \frac{\sqrt{2}}{5\sqrt{7}} & \frac{\sqrt{2}}{5\sqrt{13}} & \frac{1}{2\sqrt{65}} \\ \frac{\sqrt{3}}{2\sqrt{14}} & \frac{\sqrt{3}}{2\sqrt{35}} & \frac{3}{4} & -\frac{1}{2\sqrt{5}} & -\frac{1}{\sqrt{30}} & -\frac{\sqrt{7}}{\sqrt{390}} & -\frac{\sqrt{7}}{4\sqrt{39}} \\ \frac{\sqrt{3}}{\sqrt{70}} & \frac{\sqrt{3}}{5\sqrt{7}} & -\frac{1}{2\sqrt{5}} & \frac{4}{5} & -\frac{\sqrt{2}}{5\sqrt{3}} & -\frac{\sqrt{14}}{5\sqrt{39}} & -\frac{\sqrt{7}}{2\sqrt{195}} \\ \frac{1}{\sqrt{35}} & \frac{\sqrt{2}}{5\sqrt{7}} & -\frac{1}{\sqrt{30}} & -\frac{\sqrt{2}}{5\sqrt{3}} & \frac{2}{5} & \frac{2\sqrt{7}}{5\sqrt{13}} & \frac{\sqrt{7}}{\sqrt{130}} \\ \frac{1}{\sqrt{65}} & \frac{\sqrt{2}}{5\sqrt{13}} & -\frac{\sqrt{7}}{\sqrt{390}} & -\frac{\sqrt{14}}{5\sqrt{39}} & \frac{2\sqrt{7}}{5\sqrt{13}} & \frac{3}{5} & -\frac{1}{\sqrt{10}} \\ \frac{1}{2\sqrt{26}} & \frac{1}{2\sqrt{65}} & -\frac{\sqrt{7}}{4\sqrt{39}} & -\frac{\sqrt{7}}{2\sqrt{195}} & \frac{\sqrt{7}}{\sqrt{130}} & -\frac{1}{\sqrt{10}} & \frac{3}{4} \end{pmatrix}.$$

3 Liens avec l'échantillonnage spatial

Le théorème 2.1 et son corollaire montrent que le plan proposé aura tendance, pour une population préalablement triée, à ne pas sélectionner deux unités proches ($k - l$ petit en valeur absolue) ce qui correspond à l'esprit de l'échantillonnage spatial. La présentation s'attachera alors à illustrer par des simulations les performances des différentes stratégies.

References

Benedetti, R., Piersimoni, F., and Postiglione, P. (2015). *Sampling Spatial Units for Agricultural Surveys*. Springer.

Dickson, M. M. and Tillé, Y. (2016). Ordered spatial sampling by means of the traveling salesman problem. *Computational Statistics*, pages 1–14.

Fickus, M., Mixon, D. G., Poteet, M. J., and Strawn, N. (2013). Constructing all self-adjoint matrices with prescribed spectrum and diagonal. *Advances in Computational Mathematics*, 39(3-4):585–609.

Grafström, A. (2012). Spatially correlated poisson sampling. *Journal of Statistical Planning and Inference*, 142(1):139–147.

Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 24(2):120–131.

Horn, A. (1954). Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630.

Hough, J. B., Krishnapur, M., Peres, Y., Virág, B., et al. (2006). Determinantal processes and independence. *Probab. Surv*, 3:206–229.

Kadison, R. V. (2002). The pythagorean theorem: I. the finite case. *Proceedings of the National Academy of Sciences*, 99(7):4178–4184.

Kulesza, A. (2012). *Learning with Determinantal Point Processes*. PhD thesis, University of Pennsylvania.

Loonis, V. and Mary, X. (2015). Determinantal sampling designs. *arXiv preprint arXiv:1510.06618*.

Loonis, V. and Mary, X. (2016). Determinantal sampling designs. *Submitted*.

Macchi, O. (1975). The coincidence approach to stochastic point processes. *Advances in Applied Probability*, pages 83–122.

Schur, J. (1911). Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und Angewandte Mathematik*, 140:1–28.

Soshnikov, A. (2000). Determinantal random point fields. *Russian Mathematical Surveys*, 55(5):923–975.

Stevens Jr, D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278.