Journées de méthodologie statistique de l'Insee 2022

À la recherche du plan déterminantal optimal

Kim Antunez (*), Vincent Loonis (**)

(*) Insee, Direction de la diffusion et de l'action régionale
(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

Mots-clés. (6 maximum) : Échantillonnage déterminantal, équilibrage, échantillonnage spatial, optimisation semi-définie, recuit simulé.

Domaines. Théorie des sondages amont ; Échantillonnages particuliers.

Résumé

Loonis et Mary (2019) introduisent une nouvelle famille de plans de sondages paramétrée par les matrices hermitiennes contractantes, dont ils étudient les propriétés théoriques et pratiques : les plans déterminantaux. Ils montrent qu'une grande partie des propriétés, à distance finie ou asymptotiques, d'un plan déterminantal découle directement des caractéristiques de la matrice K à laquelle il est associé. Pour un jeu de probabilités d'inclusion simple fixé, ils identifient formellement une sous-famille de plans déterminantaux de taille fixe conduisant empiriquement à une qualité d'équilibrage sur une variable supérieure à celle observée pour d'autres méthodes « concurrentes ».

Des résultats empiriques montrent le bon comportement des plans déterminantaux dans le cas de plusieurs variables d'équilibrage ou de l'échantillonnage spatial. Ils sont cependant partiels et ne peuvent pas être associés à des plans déterminantaux connus formellement. Ils sont obtenus par des heuristiques, alors qu'une étude approfondie nécessiterait le recours à des techniques d'optimisation semi-définie non linéaire, c'est-à-dire d'optimisation dans l'ensemble des matrices hermitiennes contractantes à diagonale, et éventuellement spectre, fixés. La recherche des solutions pour de tels problèmes s'avère complexe.

Loonis (2021) propose une paramétrisation des matrices hermitiennes contractantes par un ensemble de paramètres indépendants transformant le problème d'optimisation semi-définie avec contraintes sur le spectre et la diagonale en un problème non contraint. L'application, qui à une valeur du paramètre associe une matrice hermitienne, est continue mais non différentiable pour certaines composantes, ce qui incite à mobiliser des algorithmes d'optimisation stochastiques dont le recuit-simulé (Kirkpatrick *et al.* (1983)).

Dans cette présentation, nous étudions la performance de différentes méthodes de recherche de plans déterminantaux équilibrés et spatiaux optimaux. Nous faisons varier la taille de la population ainsi que le nombre de variables auxiliaires considérées. Les résultats empiriques montrent que plus le nombre de variables auxiliaires augmente, plus la paramétrisation proposée par Loonis (2021) démontre son intérêt par rapport aux heuristiques présentées dans Loonis et Mary (2019).

Nous montrons également que l'efficacité empirique des plans de sondage déterminantaux est convaincante en comparaison d'autres méthodes d'échantillonnage de référence à probabilités d'inclusion simple inégales.

Toutefois, la nouvelle paramétrisation se confronte à la malédiction de la dimension (Bellman et Kalaba (1959)), c'est-à-dire à des difficultés qui apparaissent quand la dimension du vecteur de paramètres augmente, se traduisant en l'occurrence par une explosion du temps de calcul de l'algorithme de recuit simulé.

Abstract

Loonis et Mary (2019) introduces a new family of sampling designs parametrized by contracting Hermitian matrices: determinantal sampling designs. They obtain, by semi-definite optimization with constraints on the spectrum and the diagonal, a sub-family which leads empirically to a quality of balancing on a variable superior to the one obtained for other concurrent methods.

Loonis (2021) proposes a parametrization of contracting Hermitian matrices by a set of independent parameters transforming the optimization problem into an unconstrained problem.

In this presentation, we study the performance of these methods. We vary the size of the population as well as the number of auxiliary variables used. The empirical results show that the more the number of auxiliary variables increases, the more the new parametrization demonstrates its interest compared to the heuristics presented in Loonis et Mary (2019). However, the new parameterization is confronted with the curse of dimension (Bellman et Kalaba (1959)), that is to say with difficulties which appear when the dimension of the vector of parameters increases, resulting in this case by an explosion of the calculation time of the simulated annealing algorithm used to find the optimal determinantal matrix.

Introduction

Loonis et Mary (2019) introduisent une nouvelle et vaste famille de plans de sondages : les plans déterminantaux, paramétrés par matrices hermitiennes contractantes [1]. Les auteurs en étudient les propriétés théoriques et pratiques, notamment le fait que les probabilités d'inclusion simple et double se déduisent aisément de la matrice K. En les comparant à d'autres méthodes d'échantillonnage, ils constatent que ces plans de sondage présentent de très bonnes propriétés théoriques et empiriques d'équilibrage sur une variable, ce point étant étudié plus systématiquement dans Loonis (2021).

Une reformulation du problème d'équilibrage sur plusieurs variables conduit à l'exprimer, dans les cadre des plans déterminantaux, sous la forme d'un problème d'optimisation semi-définie non linéaire, c'est-à-dire l'optimisation dans l'ensemble des matrices hermitiennes contractantes à diagonale fixée. La recherche des solutions pour de tels problèmes s'avérant complexe, Loonis et Mary (2019) proposent dans leur article des heuristiques, s'appuyant sur le fait que la multiplication d'une matrice K par des matrices unitaires astucieusement choisies permet de conserver la diagonale et le spectre.

Loonis (2021) suggère une voie différente : une paramétrisation appropriée des matrices hermitiennes contractantes par un ensemble de paramètres indépendants permet de transformer le problème d'optimisation semi-définie en un problème non contraint dans $\mathbb{R}^m$ ou $\mathbb{C}^m$. Toutefois, le nombre de paramètres m est important, notamment quand la population étudiée est grande. Par ailleurs, l'application, qui à une valeur du paramètre associe une matrice hermitienne, est continue mais non différentiable pour certaines composantes. La non-différentiabilité incite à mobiliser des algorithmes d'optimisation stochastiques dont le recuit-simulé (Kirkpatrick *et al.* (1983)) fait partie. La très grande taille du vecteur de paramètres laisse cependant présager des difficultés liées à la malédiction de la dimension (Bellman et Kalaba (1959)), c'est-à-dire des difficultés spécifiques apparaissant quand la dimension du vecteur de paramètres augmente.

Dans cet article, nous nous intéressons donc en particulier à la performance du **recuit-simulé** pour la recherche de plans déterminantaux optimaux en terme d'équilibrage. Nous procéderons par simulation et étudierons la performance de cette méthode en fonction de la **taille de la population** et du **nombre de variables auxiliaires** considérées.

Pour les parties empiriques de cet article, nous travaillons sur les données du dataset `Meuse` (figure 1) disponible dans le package R `sp`. Cette base de données renseigne sur `N` emplacements (15mx15m) de métaux lourds situés dans une plaine inondable près du village de Stein aux Pays-Bas.

Les probabilités de tirage des individus sont inégales, proportionnelles à la variable `copper` (concentration de cuivre) et leur somme est un entier `n`.

1. 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.

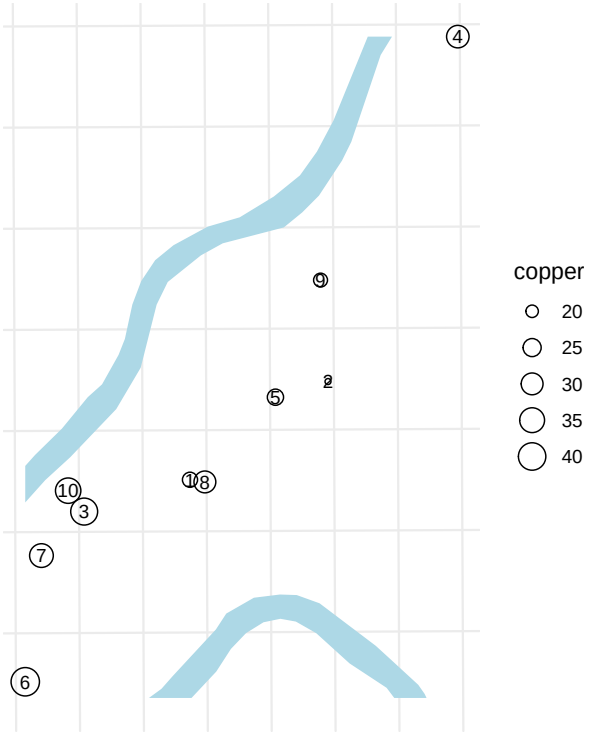

FIGURE 1 – Données `Meuse` pour `N = 10` et `n = 4`.

1 Équilibrage sur une ou plusieurs variables

Il est courant, dans la pratique et la théorie des sondages, de chercher à sélectionner des échantillons par des procédures respectant des probabilités d'inclusion simple fixées et estimant parfaitement des totaux connus *a priori* pour chaque individu de la population. Ces contraintes sont appelées contraintes d'équilibrage. Dans cet article, les contraintes d'équilibrage seront de deux types :

1. **Minimisation de la variance de Q variables auxiliaires** : Le premier type d'équilibrage visera à estimer sans biais et avec une variance la plus faible possible les totaux de Q variables auxiliaires à partir d'un échantillon aléatoire (nous testerons empiriquement $Q \in \{1,2,3\}$)[2].
2. **Échantillonnage spatial** : Le second type d'équilibrage sera équivalent à définir $Q = N$ variables auxiliaires spécifiques de manière à proposer une bonne propriété d'étalement spatial des échantillons. Cette méthode, exposée plus tard, s'inspirera de Jauslin et Tillé (2019).

Le fait de proposer ces deux méthodes nous permettra d'évaluer l'efficacité des plans de sondages déterminantaux optimaux en fonction du nombre de contraintes d'équilibrages mobilisées : faible nombre dans le cas de la variance, nombre élevé dans le cas de l'échantillonnage spatial.

2. On utilise comme variables auxiliaires les concentrations en cadmium (`cadmium`), plomb (`lead`) et zinc (`zinc`) dont on norme la somme afin qu'elles aient donc toutes la même importance quelle que soit leur unité initiale. Cela reviendra à minimiser la somme des carrés du coeffcient de variation des Q estimateurs. Ces variables sont assez fortement corrélées sans non plus être colinéaires.

1.1 Minimisation de la variance de Q variables auxiliaires

On se place tout d'abord dans le cas où l'on observe Q variables auxiliaires $x^1, \cdots, x^q, \cdots, x^Q$ et on cherche à estimer chaque total $t_{x^q} = \sum_{k \in U} x_k^q$, par son estimateur d'Horvitz-Thompson : $\hat{t}_{x^q} = \sum_{k \in S} x_k^q \pi_k^{-1}$.

L'estimateur d'Horvitz-Thompson étant sans biais, si les probabilités d'inclusion simple sont toutes strictement positives, on cherche à minimiser la fonction objectif $C(\mathscr{P})$ qui correspond à la somme des variances des estimateurs $\hat{t}_{x^q}$ (multipliée par un facteur 2 par commodité de calcul) :

$$C(\mathscr{P}) = 2 \sum_{q=1}^{Q} Var(\hat{t}_{x^q}) \tag{1}$$

Si le plan est de taille fixe, (1) devient (Särndal *et al.* (2003)) :

$$\begin{aligned} C(\mathscr{P}) &= 2\tfrac{-1}{2} \sum_{q=1}^{Q} \sum_{k=1}^{N} \sum_{l=1}^{N} \left(\frac{x_k^q}{\pi_k} - \frac{x_l^q}{\pi_l}\right)^2 (\pi_{kl} - \pi_k \pi_l) \\ &= \sum_{k=1}^{N} \sum_{l=1}^{N} \underbrace{\underbrace{\sum_{q=1}^{Q} \left(\frac{x_k^q}{\pi_k} - \frac{x_l^q}{\pi_l}\right)^2}_{Q_{kl}(x^1, \cdots, x^Q, \pi)} (\pi_k \pi_l - \pi_{kl})}_{C_{kl}(x^1, \cdots, x^Q, \pi)} \\ &= e_N^t \, C(Q, \pi) \, e_N \end{aligned} \tag{2}$$

où e_N est un vecteur de taille N dont toutes les composantes valent 1 et C est une matrice de coefficient C_{kl}.

Ainsi, pour rendre la variance petite, à probabilités d'inclusion fixées, les probabilités d'inclusions doubles π_{kl} doivent être grandes quand les individus ne se ressemblent pas au sens de $\frac{x_k^q}{\pi_k}$, et inversement.

Si l'on peut paramétriser les probabilités d'inclusion double sous la forme $\pi_{kl}(\theta)$, à probabilités d'inclusion simple fixées [3], il est alors possible de réinterpréter le problème d'équilibrage en un problème de minimisation, en θ du critère $C(\mathscr{P})$ de l'équation 2.

1.2 Échantillonnage spatial

L'échantillonnage spatial correspond à une situation où l'on dispose des coordonnées géographiques (en général en deux dimensions : longitude et latitude). Si on suppose que l'autocorrélation spatiale d'un phénomène diminue avec la distance (première loi de Tobler), il conviendra donc, à probabilités d'inclusion fixées, de privilégier la sélection d'individus distants plutôt que proches afin de sélectionner des individus suffisamment distincts dans notre échantillon. C'est pourquoi les méthodes d'échantillonnage spatial cherchent d'une part à définir un indicateur d'étalement spatial et à l'optimiser d'autre part.

Il existe différentes méthodes de tirage d'échantillons spatialement dispersés (Cube spatial, méthode GRTS, etc.). On trouvera une synthèse dans Loonis et De Bellefon (2018). Nous

3. comme c'est le cas dans les plans de sondage déterminantaux puisque $\pi_{kl} = f^{kl}(K)$, avec f^{kl} connue.

considérons ici une méthode inspirée de Jauslin et Tillé (2019). Les auteurs assimilent l'échantillonnage spatial à un plan de sondage équilibré sur $Q = N$ variables auxiliaires informant sur l'appartenance au voisinage (domaine D_q) de chaque individu spatial q.

Le domaine D_q (de cardinal d_q) correspond à l'ensemble des plus proches voisins de l'unité q (incluant q). On ajoute les voisins un à un (dans leur ordre de proximité à l'unité k) à la liste des plus proches voisins jusqu'à ce que la somme des probabilités d'inclusion simple soit supérieure ou égale à 1. La figure 2 illustre les $Q = N$ domaines des données `Meuse` pour les probabilités d'inclusion retenues.

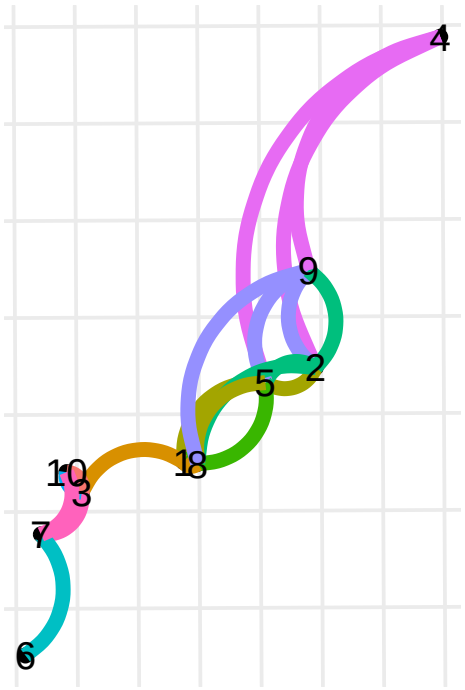

FIGURE 2 – Domaines D_q pour `N = 10`.

Note : Chaque domaine q correspond à l'ensemble des entités géographiques reliées à q par une courbe.

En prenant comme variable auxiliaire $x_k^q = \pi_k 1(k \in D_q)$, il vient $t_{x_q} = \sum_{k \in U} \pi_k 1(k \in D_q) = \sum_{k \in D^q} \pi_k \simeq 1$, par construction. Par ailleurs, l'estimateur d'Horvitz-Thompson de t_{x_q} est $\hat{t}_{x_q} = \sum_{k \in \mathcal{S}} (\pi_k 1(k \in D_q))/\pi_k = \sharp \mathcal{S} \cap D_q$, taille de l'échantillon dans le domaine D_q. Équilibrer sur x^q, revient donc chercher à avoir 1 individu et 1 seul échantillonné dans le domaine D_q. Le critère à optimiser se met alors sous la forme :

$$
\begin{aligned}
C(\mathscr{P}) &= \sum_{k=1}^{N}\sum_{l=1}^{N}\sum_{q=1}^{Q}\left(\frac{x_k^q}{\pi_k}-\frac{x_l^q}{\pi_l}\right)^2(\pi_k\pi_l-\pi_{kl}) \\
&= \underbrace{\sum_{k=1}^{N}\sum_{l=1}^{N}\underbrace{\sum_{q=1}^{Q}\left(1(k\in D_q)-1(l\in D_q)\right)^2}_{Q_{kl}(x^1,\cdots,x^Q,\pi)}(\pi_k\pi_l-\pi_{kl})}_{C_{kl}(x^1,\cdots,x^Q,\pi)} \\
&= e_N^t\, C(Q,\pi)\, e_N
\end{aligned}
\tag{3}
$$

Le terme Q_{kl} est le nombre de domaines dans lesquels il y a un des deux individus k ou l mais pas les deux. Si deux individus sont très proches géographiquement, ce terme aura tendance à être faible. S'ils sont éloignés, il sera grand (voir figure 3). Avec le même raisonnement que précédemment, minimiser le critère C revient à privilégier le tirage d'individus éloignés .

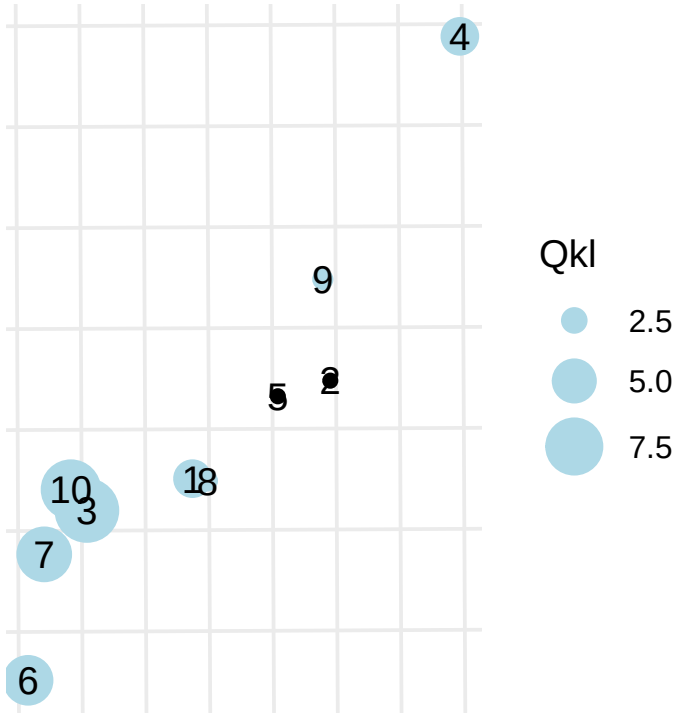

FIGURE 3 – Un exemple de la valeur de la matrice $Q_{k,l}$ pour $k = 5$ (`N = 10`).

Ainsi, l'objectif d'étalement spatial de l'échantillon quand il se formule à la manière d'un problème d'équilibrage sur $Q = N$ variables, se traduit par un problème de minimisation du critère $C(\mathscr{P})$ de l'équation 3. Dans la suite, nous appellerons « critère géographique » la valeur de la fonction objectif, pour un plan donné, dans le cas $Q = N$ et les variables auxiliaires définies ici.

1.3 Lien avec les plans déterminantaux

Compte tenu des définitions des plans déterminantaux présentées (voir encadré 1), si le plan est déterminantal associé à une matrice hermitienne contractante K et est de taille fixe, les équations

2 et 3 se mettent sous une forme qui est directement fonction de K :

$$\begin{aligned} C(DSD(K)) &= \sum_{k=1}^{N}\sum_{l=1}^{N} Q_{kl}(x^1,\cdots,x^Q,diag(K))\,|K_{kl}|^2 \\ &= e_N^t \underbrace{Q * K * \overline{K}}_{\text{C(Q,K)}} e_N \\ &= e_N^t\, C(Q,K)\, e_N \end{aligned} \tag{4}$$

où $*$ désigne le produit matriciel de Hadamard (multiplication terme à terme).

Encadré 1 : Définitions des plans de sondage déterminantaux
Une variable aléatoire $\mathbb{S}$ à valeurs dans l'ensemble des parties de U, $\mathcal{P}(U)$, a pour loi de probabilité un plan de sondage déterminantal si et seulement si il existe une matrice hermitienne contractante K indexée par U, appelée noyau, telle que pour tout $s \in \mathcal{P}(U)$,

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

où $K_{|s}$ est la sous matrice de K indéxée par les unités de s. On utilisera alors la notation $\mathbb{S} \sim DSD(K)$. Il est alors possible de paramétriser un plan de sondage sous la forme d'une matrice appelée noyau appartenant à la large famille des matrices hermitiennes contractantes.
De cette définition découle directement le calcul des probabilités d'inclusion simple et double à partir de la matrice K : les probabilités d'inclusion simple apparaissent sur sa diagonale et les probabilités d'inclusion double s'obtiennent, à probabilités d'inclusion données, à partir des termes hors diagonale (propriété 1.1).

Propriété 1.1 (Probabilités d'inclusion et matrice K)
$\pi_k = pr(k \in \mathbb{S}) = \det(K_{|\{k\}}) = K_{kk}$

$$\pi_{kl} = pr(k,l \in \mathbb{S}) = \det\begin{pmatrix} K_{kk} & K_{kl} \\ \overline{K}_{kl} & K_{ll} \end{pmatrix} = K_{kk}K_{ll} - |\,K_{kl}\,|^2 \qquad (k \neq l)$$

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l = -|\,K_{kl}\,|^2 & (k \neq l) \\ \pi_k(1-\pi_k) = K_{kk}(1-K_{kk}) & (k = l) \end{cases} \tag{5}$$

2 Trois stratégies d'optimisation des plans de sondages déterminantaux

Trouver un plan de sondage déterminantal optimal revient donc à trouver la matrice K qui minimise le critère $C(Q,K)$ de l'équation 4.

Toute la difficulté réside dans la recherche d'une méthode permettant d'optimiser d'un tel critère en parcourant toutes les matrices K candidates. L'utilisation d'algorithmes d'optimisation semi-définie positive [4], respectant les contraintes de trouver une matrice, à la fois hermitienne et à la diagonale fixée, est un problème difficile. Quand on souhaite, en outre, obtenir un plan de taille fixe, il faut nécessairement que la matrice obtenue soit une projection et donc de spectre fixé.

4. L'optimisation semi-définie positive (ou SDP) est un type d'optimisation où l'inconnue est une matrice hermitienne dont les valeurs propres sont positives.

Trois pistes de recherche d'une matrice K minimisant les critères d'équilibrages sont comparées dans la suite de cet article. Elles sont ordonnées de la méthode la moins à la plus aboutie mais également de la moins chronophage à la plus chronophage computationnellement :

1. **Les bonnes propriétés de la matrice P^{Π} triée astucieusement** : Loonis et Mary (2019) montrent qu'une matrice de projection particulière appelée P^{Π} présente des propriétés intéressantes de répulsion (exclure au maximum de tirer des unités qui se ressemblent) quand les individus sont triés de manière à ce que deux individus dont l'ordre d'apparition est proche soient deux individus qui se ressemblent.
2. **Améliorer P^{Π} par rotations successives** : Une modification itérative d'une matrice P^{Π} est rendue possible par des heuristiques. Elles exploitent le fait que la composition d'une matrice K par des matrices de rotation astucieusement choisies permet d'obtenir une nouvelle matrice hermitienne ayant la même diagonale et le même spectre que la matrice d'origine.
3. **La paramétrisation des matrices K avec un grand nombre de paramètres** : Loonis (2021) propose une paramétrisation des matrices hermitiennes contractantes par un ensemble de $2Nn$ paramètres indépendants transformant le problème d'optimisation semi-définie en un problème non contraint mais comportant de nombreux paramètres.

2.1 Les bonnes propriétés de la matrice P^{Π} triée astucieusement

Loonis et Mary (2019) exhibent une matrice de projection particulière – la matrice P^{Π} dont les coefficients sont donnés en encadré 2 – qui présente des propriétés intéressantes en termes de répulsion.

Encadré 2 : Les coefficients de la matrice P^{Π}

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

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

— $1 < k_r \leq N$ est un entier tel que $\sum_{k=1}^{k_r-1} \Pi_k < r$ et $\sum_{k=1}^{k_r} \Pi_k \geq r$; par convention on posera $k_0 = 0$

— $\alpha_{k_r} = r - \sum_{k=1}^{k_r-1} \Pi_k$. On notera que $\alpha_{k_r} = \Pi_{k_r}$ si $\sum_{k=1}^{k_r} \Pi_k = r$.

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

On peut montrer que les probabilités d'inclusion double de la matrice P^{Π}, à tri des individus fixé, sont nulles si k et l ont des ordres de classement proches, et maximales si k et l ont des ordres éloignés. Pour répondre à l'objectif de minimisation des critères mis en évidence dans les

équations 2 et 3, il devient alors intéressant de trier les données de telle sorte que deux individus dont l'ordre d'apparition est proche soient deux individus qui se ressemblent :

— Lorsque $Q = 1$, cela revient à trier les individus selon $\frac{x_k}{\pi_k}$ puisque les individus proches ont un $\frac{x_k}{\pi_k}$ proche. lorsque $Q > 1$, il n'y a pas de manière unique de trier les individus : plusieurs stratégies de tri multi-dimensionnel peuvent être pertinentes sans relation d'ordre total entre elles. Nous choisissons de trier les individus selon la somme $\sum_{q=1}^{Q} x_k^q / \pi_k$.

— Pour minimiser le critère géographique (équation 3), les individus qui se ressemblent sont ceux qui sont proches géographiquement. On peut alors trier les individus selon un chemin de Hamilton, c'est-à-dire dessiner le chemin le plus court qui relie l'ensemble des points (figure 4).

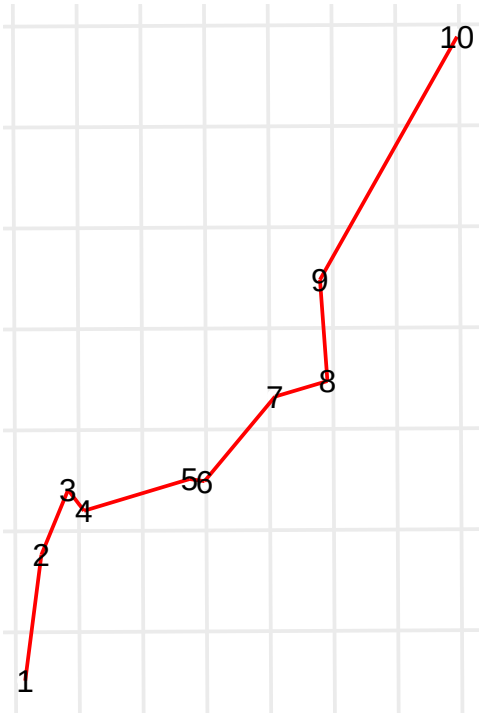

FIGURE 4 – Chemin de Hamilton sur données `Meuse` pour `N = 10`.

Note : L'algorithme de calcul du chemin d'Hamilton est la méthode `two_opt` de la fonction `solve_TSP` du package R `TSP`, une heuristique (Croes (1958)) dont le principe est d'échanger deux arêtes dans le graphe jusqu'à ce qu'aucune modification ne soit possible.

2.2 Améliorer P^{Π} par rotations successives

Si K est une matrice telle que $K_{kk} \neq K_{ll}$ pour un couple d'individus, k et l, il existe une matrice unitaire $R_{kl}(\theta_{kl})$, construite sur la base d'une rotation dont les auteurs exhibent la valeur spécifique de l'angle θ_{kl}, telle que le produit $R_{kl}(\theta_{kl})KR_{kl}^T(\theta_{kl})$ a la même diagonale et le même spectre que K.

Comme P^{Π} présente de bonnes propriétés, il peut être intéressant de partir de la valeur initiale $K^0 = P^{\Pi}$, construite sur une population astucieusement triée, puis de modifier itérativement cette

matrice de manière à faire diminuer au maximum le critère $C(DSD(K))$ (équation 4) à minimiser. Cette démarche est réalisée via l'algorithme 2.1 ci-après [5]. On suppose ici que $\Pi_k \neq \Pi_l$ pour tout couple k,l.

Algorithme 2.1 (Algorithme de minimisation de $C(DSD(K))$)

1. *Initialiser l'algorithme à* $K = K^0 = P^{\Pi}$.
2. *For couple* $= 1$ *à* $\frac{N(N-1)}{2}$ *:*
   (a) *Tirer au hasard un couple* (k,l) *tel que* $l < k$ *non tiré jusqu'à présent.*
   (b) *Calculer* $K' = R_{kl}(\theta_{kl})KR_{kl}^T(\theta_{kl})$.
   (c) *Si* $C(DSD(K')) < C(DSD(K))$ *alors* $K \leftarrow K'$.
3. *Fin de for.*
4. *Réitérer l'étape 2. jusqu'à ne plus trouver de rotation qui diminuent le critère.*
5. *Retourner K.*

2.3 La paramétrisation des matrices K avec un grand nombre de paramètres

En s'inspirant de la méthode de Fickus *et al.* (2013) permettant de construire toutes les matrices hermitiennes de diagonale et de spectre fixés, Loonis (2021) propose une nouvelle paramétrisation des plans de sondage déterminantaux à partir de deux matrices de grandes dimensions. Dans le cas des plans de taille fixe, il s'agit de :

— $\Omega[N,n]$ dont la k^{me} colonne conditionne la loi de la taille de l'échantillon dans le domaine des points de 1 à k. Tous les coefficients de $\Omega[N,n]$ sont indépendants et à valeur dans $[0,1]$.

— $\rho[N,n]$ qui, à variance de la taille de l'échantillon dans chacun des domaines 1 à k fixée par Ω, conditionne les termes hors diagonal de K, et donc les probabilités d'inclusion double. Tous les coefficients de $\rho[N,n]$ sont indépendants et à valeur dans $[0,1]$.

Les plans de sondages déterminantaux deviennent alors complètement paramétrisés par des paramètres indépendants.

Toutefois, le nombre de paramètres ($2nN$) est important et explose quand la population étudiée devient grande.

Enfin, l'application qui, à une valeur du paramètre associe une matrice hermitienne, est bien continue mais est non différentiable pour certaines composantes. L'aspect non différentiable incite donc à mobiliser des algorithmes d'optimisation stochastiques, dont le recuit-simulé fait partie. Celui-ci est mis en œuvre en partie 3. La très grande taille du vecteur de paramètres laisse cependant présager des difficultés liées à la malédiction de la dimension (Bellman et Kalaba (1959)), c'est-à-dire des difficultés spécifiques apparaissant quand la dimension du vecteur de paramètres augmente.

5. Cet algorithme n'est pas le seul possible. Nous pourrions également imaginer des variantes comme la multiplication par des rotations sans fixer l'angle mais en ajoutant une contrainte non linéaire pour fixer la diagonale, ou encore la multiplication non pas par une mais par plusieurs rotations.

3 Optimisation des paramètres par recuit simulé

3.1 Recuit simulé

Le recuit simulé (Kirkpatrick *et al.* (1983)) est une méthode empirique d'optimisation permettant de trouver les extrema d'une fonction.

Il s'appuie sur l'algorithme de Metropolis-Hastings. Nous en faisons ici une utilisation légèrement simplifiée [6] (pseudo-code de l'algorithme 3.1). On part d'un « état » donné du système. Dans notre cas, il s'agit d'une valeur initiale des paramètres Ω et ρ prise au hasard. À chaque itération de l'algorithme, on modifie l'état pour en obtenir un autre, voisin. Le voisin correspond à l'état précédent auquel on ajoute la génération d'une loi uniforme sur $[-p;p]$ avec p un pas choisi [7].

Soit l'état voisin améliore le critère que l'on cherche à optimiser, soit celui-ci le dégrade. On accepte un état s'il améliore le critère pour ainsi tendre à trouver l'optimum dans le voisinage de l'état de départ.

Algorithme 3.1 (Pseudo-code de l'algorithme de recuit simulé simplifié)

```
Fonction recuit_simule(s0, C, niter)
  s := s0   #état initial
  c := C(s) #critère initial
  k := 0  #numéro de l'itération
  pour k in 1:niter
    sn := voisin(s, pas)
    cn := C(sn)
    si en < e alors #condition à tester
      s := sn; e := en
    k := k + 1
  retourne s
Fin Fonction
```

Cet algorithme de recuit simulé est encapsulé dans l'algorithme 3.2 de recherche de la matrice K optimale.

Algorithme 3.2 (Pseudo-code de l'algorithme de recherche de K optimal par recuit simulé)

```
Nb_tirages_initiaux = 100 000
Nb_tirages_gardes = 10
niter = 400 000

Fonction recuit_simule()
```

6. Dans une version plus complète, on peut également permettre l'acceptation d'un « mauvais » état avec une certaine probabilité, afin d'explorer une plus grande partie de l'espace des états et d'éviter de s'enfermer trop vite dans la recherche d'un optimum local. On remplace alors la condition à tester par `cn < c ou runif(1) < exp(-(c_n - c_c) / (1 - constante)^k)` (règle de Métropolis) avec `runif(1)` qui envoie une valeur aléatoire dans l'intervalle $[0;1]$. Il est également possible d'ajouter un critère d'arrêt tel que le fait de continuer jusqu'à un nombre maximal d'itération (`k < kmax`) ou encore s'arrêter quand un état dont le critère est minimal est trouvé (`e < emin`). Nous n'avons pas encore testé ces éléments de raffinement au moment de la rédaction de l'article.

7. Nous choisissons $p = 0,01$ dans nos différentes simulations (meilleure efficacité empirique après quelques tests). Nous pourrions également faire diminuer ce pas progressivement au fur et à mesure des itérations.

```
    pour i in 1:Nb_tirages_initiaux
      liste_s[i] = runif(2nN) #paramètres Omega et rho
      liste_c[i] = C(liste_s[i]) #Critere appliqué aux paramètres
  indices = indices des Nb_tirages_gardes meilleurs éléments de liste_c
  liste_s_gardes = liste_s[indices]
  pour i in 1:Nb_tirages_gardes
    liste_s_finaux[i] = recuit_simule(liste_s_gardes[i], C, niter)
  return liste_s_finaux
Fin Fonction
```

Les étapes indiquées en bleu et rouge peuvent être parallélisées. La parallélisation de l'étape en bleu n'apporte pas un gain de temps significatif puisqu'il s'agit d'une multitude de petites opérations rapides et le temps d'envoi du calcul aux différents cœurs de la machine peut être plus coûteux qu'avantageux. En revanche, la parallélisation de l'étape en rouge peut faire gagner un temps significatif.

3.2 Résultats

3.2.1 Comparaison des trois méthodes d'optimisation des plans déterminantaux

Dans cette dernière partie 3.2, nous comparons les performances des trois stratégies d'optimisation des plans de sondage déterminantaux exposées en partie 2, à savoir la construction de la matrice P^{Π} (partie 2.1) construite à partir d'une population astucieusement triée, son optimisation par rotations successives (partie 2.2) et la paramétrisation totale des plans de sondages par Ω et ρ avec optimisation du critère par recuit simulé (parties 2.3 et 3.1).

Nous évaluerons la performance des trois méthodes en fonction du nombre de variables auxiliaires considérées : de une à trois en ce qui concerne le critère de la somme de Q variables explicatives (partie 1.1) et un équivalent de N variables auxiliaires pour le critère géographique (partie 1.2).

Pour `N = 10`, l'algorithme 3.2 permet d'obtenir les résultats illustrés dans le tableau 1.

	Variance d'1 variable	Variances de 2 variables	Variances de 3 variables	Critère géographique (Q=N)
Matrice P^{Π}	1.757	3.604	5.154	4.515
Matrice P^{Π} puis rotations	1.757	3.604	5.097	4.515
Matrice par recuit simulé	1.757	3.591	5.073	4.249

TABLEAU 1 – Comparaison des 3 critères pour les 3 méthodes d'optimisation des plans déterminantaux (`N = 10`).

Pour le critère de minimisation de la variance d'une seule variable auxiliaire, P^{Π} trié selon l'ordre des $\frac{x_k}{\Pi_k}$ semble correspondre au minimum global puisque ni l'optimisation par composition de P^{Π} par des rotations, ni la paramétrisation par Ω et ρ suivie d'une optimisation par recuit simulé ne permet d'améliorer le critère. Ce résultat empirique illustre en réalité la conjecture de Loonis

(2021), selon laquelle le plan associé à P^{Π} serait un plan optimal. On remarque par ailleurs que les probabilités d'inclusion doubles obtenues à l'issue du recuit simulé sont, aux erreurs d'arrondis des algorithmes près, identiques à celles obtenues avec P^{Π} (figure 5).

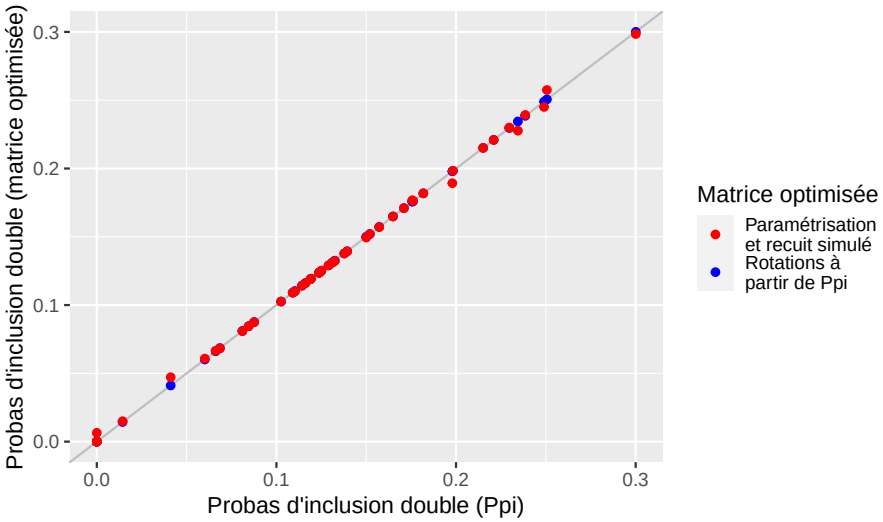

FIGURE 5 – Comparaison des probabilités d'inclusion double (critère de variance d'une variable auxiliaire) pour `N = 10`.

Plus le nombre de variables auxiliaires augmente, plus les deux méthodes d'optimisation démontrent leur performance. La composition par des rotations (partie 2.2) est computationnellement très rapide mais ne permet pas de balayer tout le spectre des plans de sondages déterminantaux. La paramétrisation (partie 2.3) suivie d'un recuit, plus chronophage mais balayant l'ensemble des plans de sondages possible, permet d'obtenir une meilleure minimisation du critère. On remarque que les probabilités d'inclusion double des deux matrices optimisées obtenues s'éloignent significativement de celles de P^{Π}, pour le tri considéré, mais sont assez proches entre elles (figure 6), ce qui laisse penser qu'on tend vers un minimum global.

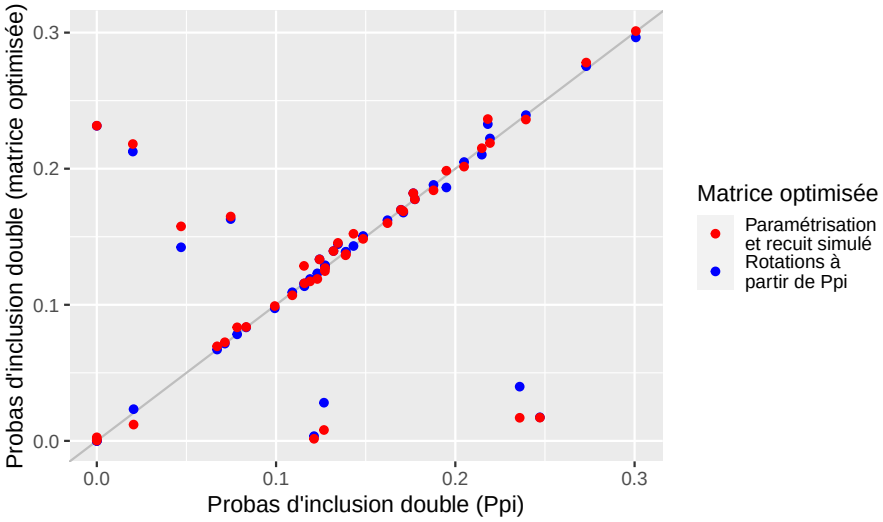

FIGURE 6 – Comparaison des probabilités d'inclusion double (Critère géographique : Q=N) pour `N = 10`.

3.2.2 Comparaison avec d'autres méthodes d'échantillonnage

Nous avons ensuite mesuré l'efficacité empirique des plans de sondage déterminantaux en les comparant à d'autres méthodes d'échantillonnage de référence à probabilités d'inclusion fixées :

— des **méthodes aléatoires** : le « Tirage systématique trié aléatoirement » (Madow (1949)) et l'échantillonnage par « Maximum d'entropie ou conditionnel de poisson » (Chen et Liu (1997)) ;
— des **méthodes d'équilibrage** : le « Tirage poissonien équilibré » (Grafström (2010)), le « Pivot équilibré » (Deville et Tillé (1998)), et le « CUBE équilibré » (Deville et Tillé (2004)) ;
— des **méthodes maximisant l'étalement spatial** : le « Tirage poissonien corrélé spatialement » (Bondesson et Thorburn (2008)), le « Pivot spatial 1 » (Grafström (2012)), le « CUBE spatial » (Grafström et Tillé (2013)) et le « WAVE sampling » (Jauslin et Tillé (2019)).

Pour chacune de ces méthodes, nous avons tiré 10 000 échantillons [8]. Nous avons ensuite calculé les quatre critères d'équilibrage du tableau 1 ainsi que deux autres indicateurs d'étalement spatial des échantillons :

— un indicateur de Moran I_{B_1} modifié pour être bornés entre -1 et 1 (Tillé *et al.* (2018)), une valeur proche de -1 indique une bonne dispersion spatiale ;
— un indicateur de l'équilibre spatial des polygones de Voronoï B (Stevens Jr et Olsen (2004)), une valeur proche de 0 indique une bonne dispersion spatiale .

Les résultats sont donnés en tableau 2 pour le critère de variance et tableau 3 pour le critère géographique et les deux critères d'étalement spatial.

Les performances des plans de sondages déterminantaux sont bonnes en comparaison des nombreuses méthodes « concurrentes » présentées. En effet, Loonis et Mary (2019) avaient déjà constaté dans le cadre d'une application sur données réelles [9] que le plan de sondage déterminantal présente de meilleures performances que ses équivalents non déterminantaux (en particulier la méthode du Cube), en termes de minimisation de la somme des variances d'une voire deux variables auxiliaires, même si le gain est parfois faible.

8. Les tirages des échantillons des plans de sondage déterminantaux s'effectuent grâce à l'algorithme 2.1 de Loonis et Mary (2019).

9. L'application de l'article concerne la sélection par l'Insee des logements à enquêter pour ses enquêtes auprès des ménages. Ce tirage s'effectue selon le principe d'un échantillon maître à deux degrés. Les auteurs ont considéré une version simplifiée à un degré qui consiste à tirer les unités primaires (UP, regroupement d'entités géographiques contiguës) dont les probabilités d'inclusion sont proportionnelles au nombre d'habitants. Les deux variables d'équilibrage sont le montant total des allocations de chômage et des revenus imposables.

	var1	var2	var3
Systematique aléatoire	10.242	14.792	18.956
Maximum d'entropie	10.443	15.053	19.258
Poisson équilibré	4.125	5.607	6.801
Pivot équilibré 1	2.995	4.299	5.665
Cube équilibré	2.811	4.733	6.572
Déterminantal (P^{Π})	1.757	3.604	5.154
Déterminantal (P^{Π} puis rotations)	1.757	3.604	5.097
Déterminantal (Recuit simulé)	1.757	3.591	5.073

TABLEAU 2 – Critères de minimisation de la somme des variances (`N = 10`).

[a] Sont ici représentés la moyenne sur 10 000 échantillons, sauf pour l'échantillonnage déterminantal pour lequel le critère théorique du tableau 1 est reporté. Pour représenter un échantillon équilibré, Le critère de la somme des variances (sur 1, 2, ou 3 variables) doit être proche de 0.

	IB1	B	geo
Systematique aléatoire	-0.175	0.299	10.514
Maximum d'entropie	-0.162	0.301	10.671
Poisson local	-0.432	0.189	5.718
Pivot local 1	-0.491	0.168	4.633
Cube local	-0.441	0.184	5.490
Wave Sampling	-0.569	0.164	3.354
Déterminantal (P^{Π})	-0.264	0.278	4.515
Déterminantal (P^{Π} puis rotations)	-0.526	0.147	4.515
Déterminantal (Recuit simulé)	-0.541	0.149	4.249

TABLEAU 3 – Critères géographique et d'étalement spatial (`N = 10`).

[a] Sont ici représentés la moyenne sur 10 000 échantillons, sauf pour l'échantillonnage déterminantal, critère géographique, pour lequel le critère théorique du tableau 1 est reporté. Pour représenter un échantillon spatialement étendu, IB1, l'indice de Moran borné, doit être proche de -1 alors que B, l'indicateur de Voronoï, et geo, le critère géographique, doivent être proches de 0.

Même si l'optimisation des plans de sondages déterminantaux fait diminuer le critère initial de P^{Π}, en particulier quand le nombre de variables auxiliaires est élevé (cadre du critère géographique), une méthode reste meilleure concernant le critère géographique : le *Wave Sampling*.

Comme d'autres méthodes d'échantillonnage spatial, comme celle du pivot local, les plans déterminantaux vérifient les conditions de Sen-Yates-Grundy ($\pi_{kl} \leq \pi_k \pi_l \quad (k \neq l)$). Cela assure la positivité de l'estimateur de la variance mais a pour désavantage d'avoir moins de flexibilité concernant la valeur local du critère C qui est nécessairement positif. A l'inverse, le *Wave*

Sampling, tout comme la méthode du Cube, ne respectent pas nécessairement ces conditions et peuvent ainsi proposer un plus large spectre de combinaisons de probabilités d'inclusion double.

Enfin, quand on double la taille de la population, c'est-à-dire pour `N = 20`, la malédiction de la dimension (Bellman et Kalaba (1959)) fait que les paramètres utilisés dans l'algorithme 3.2 (`Nb_tirages_initiaux = 100 000`, `Nb_tirages_gardes = 10` et `niter = 400 000`) ne suffisent pas à converger vers des solutions aussi satisfaisantes que pour `N = 10`.

Il serait donc nécessaire d'augmenter le nombre d'itérations de l'algorithme.

Conclusion

Dans cet article, nous avons entrepris une première étude systématique et empirique des propriétés d'optimalité des plans déterminantaux. Nous avons utilisé une paramétrisation maniable de ces derniers, héritée de Fickus *et al.* (2013) et adaptée au cadre de la théorie des sondages, pour mobiliser des méthodes d'optimisation semi-définies relativement simples.

Sur des tailles de population faibles, nous ne sommes pas parvenus à remettre en cause la conjecture de Loonis (2021) relative à la borne inférieure de l'estimateur d'Horvitz-Thompson. Nous avons également observé un bon comportement des plans déterminantaux pour l'équilibrage sur un nombre réduit de variables. Le respect des conditions de Sen-Yates-Grundy peut cependant s'avérer contraignant quand il s'agit d'équilibrer sur un grand nombre de variables, comme c'est le cas dans l'échantillonnage spatial. Une approche telle que celle de Jauslin et Tillé (2019) est alors plus intéressante dans ce cas.

Des pistes d'approfondissement de nos travaux existent. Il s'agirait d'étudier l'apport du calcul parallèle pour traiter des tailles de population plus importantes, et donc réalistes, avec le recuit simulé. Une autre piste réside dans l'identification des paramètres les plus influents par l'analyse de sensibilité et les indices de Sobol (Sobol (2001)). Des travaux récents sur l'optimisation sur les variétés (Absil *et al.* (2009), Boumal *et al.* (2014)) semblent également prometteurs alors que ceux s'appuyant sur des techniques de lagrangien augmenté pourraient être également étudiés (Fiala *et al.* (2013)).

4 Références

Absil, P.-A., Mahony, R. et Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Bellman, R. et Kalaba, R. (1959). On adaptive control processes. *IRE Transactions on Automatic Control.*, 4(2):1–9.

Bondesson, L. et Thorburn, D. (2008). *A List Sequential Sampling Method Suitablen for Real-Time Sampling.* Scandinavian Journal of Statistics 35.3, p. 466–483.

Boumal, N., Mishra, B., Absil, P.-A. et Sepulchre, R. (2014). Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459.

Chen, S. et Liu, J. (1997). Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, 7:875–892.

CROES, G. (1958). *A method for solving traveling-salesman problems.* Operations Research, 6(6):791–812.

DEVILLE, J.-C. et TILLÉ, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912.

DEVILLE, J.-C. et TILLÉ, Y. (1998). *Unequal probability sampling without replacement through a splitting method.* Biometrika 85.1, p. 89–101.

FIALA, J., KOČVARA, M. et STINGL, M. (2013). Penlab: A matlab solver for nonlinear semidefinite optimization. *arXiv preprint arXiv:1311.5240*.

FICKUS, M., MIXON, D. G., POTEET, M. J. et STRAWN, N. (2013). Constructing all self-adjoint matrices with prescribed spectrum and diagonal. *Advances in Computational Mathematics*, 39(3-4):585–609.

GRAFSTRÖM, A. (2010). On a generalization of poisson sampling. *Journal of Statistical Planning and Inference*, 140(4)::982–991.

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

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

JAUSLIN, R. et TILLÉ, Y. (2019). Spatial spread sampling using weakly associated vectors. *arXiv preprint arXiv:1910.13152*.

KIRKPATRICK, S., GELATT, C. D. et VECCHI, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

LOONIS, V. (2021). Construire tous les plans de sondage déterminantaux. *Colloque francophone sur les sondages, Bruxelles, article soumis, DOI: 10.13140/RG.2.2.21214.92483.*

LOONIS, V. et DE BELLEFON, M.-P. (2018). *Manuel d'analyse spatiale. Théorie et mise en œuvre pratique avec R, Insee Méthodes n 131, Insee, Eurostat, 392 p.* Insee.

LOONIS, V. et MARY, X. (2019). Determinantal sampling designs. *Journal of Statistical Planning and Inference*, 199:60–88.

MADOW, W. G. (1949). On the theory of systematic sampling, ii. *The Annals of Mathematical Statistics*, pages 333–354.

SÄRNDAL, C.-E., SWENSSON, B. et WRETMAN, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

SOBOL, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.

STEVENS JR, D. L. et OLSEN, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278.

TILLÉ, Y., DICKSON, M. M., ESPA, G. et GIULIANI, D. (2018). Measuring the spatial balance of a sample: A new measure based on the moran's i index. *Spatial Statistics*, 23:182–192.