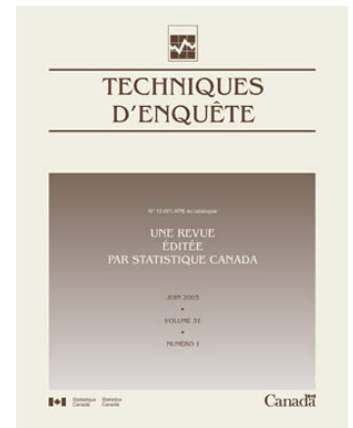


## Techniques d'enquête

# Échantillonnage graphique de population finie

par Bardia Panahbehagh

Date de diffusion : le 29 juin 2026



Statistique  
Canada

Statistics  
Canada

Canada

---

## Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca).

Vous pouvez également communiquer avec nous par :

**Courriel** à [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)

**Téléphone** entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- |                                                                             |                |
|-----------------------------------------------------------------------------|----------------|
| • Service de renseignements statistiques                                    | 1-800-263-1136 |
| • Service national d'appareils de télécommunications pour les malentendants | 1-800-363-7629 |
| • Télécopieur                                                               | 1-514-283-9350 |

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site [www.statcan.gc.ca](http://www.statcan.gc.ca) sous « Contactez-nous » > « [Normes de service à la clientèle](#) ».

## Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Publication autorisée par la ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par la ministre de l'Industrie, 2026

L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

*This publication is also available in English.*

---

# Échantillonnage graphique de population finie

Bardia Panahbehagh<sup>1</sup>

## Résumé

La présente étude propose une méthode novatrice et intuitive d'échantillonnage de population finie, élaborée à partir d'un cadre graphique original. Dans cette approche, les probabilités d'inclusion de premier ordre sont représentées sous forme de barres sur un graphique bidimensionnel. En changeant la position de ces barres, les chercheurs peuvent générer un large éventail de plans d'échantillonnage différents. Cette visualisation graphique des plans d'échantillonnage facilite l'exploration d'autres plans et peut simplifier certains aspects de la mise en œuvre par rapport aux algorithmes mathématiques traditionnels. Cette approche novatrice présente un potentiel important pour relever des défis complexes en matière d'échantillonnage, comme l'atteinte d'un plan optimal. L'application d'une version de l'algorithme de recherche glouton du meilleur à prime abord à cette approche graphique illustre le potentiel d'intégration d'algorithmes intelligents à l'échantillonnage de population finie.

**Mots-clés :** Algorithme intelligent; plan de sondage; plan optimal; probabilités d'inclusion.

## 1. Introduction

Soit  $U = \{1, \dots, N\}$ , une population de taille  $N$ , et

$$\mathcal{S} = \{s_1, s_2, \dots, s_T\},$$

l'ensemble de tous les échantillons ou sous-ensembles possibles de  $U$ . L'objectif est de sélectionner un échantillon aléatoire  $\mathbb{S}$  à partir de  $\mathcal{S}$  avec des probabilités d'inclusion de premier ordre (PIPO)  $\pi = \{\pi_k, k \in U\}$ , de taille  $n_{\mathbb{S}}$  où

$$\sum_{k \in U} \pi_k = E(n_{\mathbb{S}}).$$

Soit  $\mathbf{p} = \{p_t = \Pr(\mathbb{S} = s_t); t = 1, 2, \dots, T\}$ , un plan de sondage attribuant une probabilité à chaque échantillon possible, de sorte que :

$$p_t > 0, \sum_{t=1}^T p_t = 1 \quad \text{et} \quad \pi_k = \sum_{t: s_t \ni k} p_t \quad \text{pour toute} \quad k \in U. \quad (1.1)$$

Les probabilités d'inclusion de deuxième ordre (PIDO) peuvent être représentées sous forme matricielle comme suit :

$$\Pi = [\pi_{k\ell}]_{N \times N}, \quad \text{où} \quad \pi_{k\ell} = \sum_{t: s_t \ni k, \ell} p_t, \quad k, \ell \in U.$$

De plus, l'entropie du plan  $\mathbf{p}$  est définie par

$$H(\mathbf{p}) = - \sum_{t=1}^T p_t \ln(p_t).$$

Considérons maintenant  $y_k, x_k, k \in U$ , respectivement, comme variables principale et auxiliaire. L'objectif principal de l'échantillonnage est généralement d'estimer des paramètres comme le total de la population, désigné par

1. Bardia Panahbehagh, Faculty of Mathematical Sciences and Computer, Kharazmi University, Téhéran, Iran. Courriel : panahbehagh@khu.ac.ir.

$$Y = \sum_{k \in U} y_k.$$

L'estimation des paramètres de la population, comme un total, ainsi que l'estimation de la variance des estimateurs des paramètres reposent sur le plan de sondage, plus précisément les PIPO et les PIDO. Étant donné que  $\pi_k > 0$  pour toute  $k \in U$ , un estimateur sans biais de  $Y$  est l'estimateur Narain-Horvitz-Thompson (NHT; Narain, 1951; Horvitz et Thompson, 1952)

$$\hat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k},$$

avec

$$\text{var}(\hat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} (\pi_{k\ell} - \pi_k \pi_\ell). \quad (1.2)$$

En supposant que toutes les  $\pi_{k\ell}$  sont strictement supérieures à zéro, un estimateur sans biais de cette variance est

$$\widehat{\text{var}}(\hat{Y}) = \sum_{k \in S} \sum_{\ell \in S} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell} \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}}. \quad (1.3)$$

Cette approche traditionnelle fait appel à des plans fondés sur des algorithmes théoriques et mathématiques qui, bien que solidement établis, peuvent poser des difficultés dans certaines situations. En particulier, l'imposition de contraintes ou d'exigences supplémentaires dans ces plans conventionnels entraîne souvent une complexité accrue. Cette complexité peut se traduire par une difficulté de mise en œuvre des plans contraints avec un ensemble donné de PIPO ou par une plus grande difficulté à calculer les PIDO correspondantes, qui sont toutes deux essentielles à une estimation et à une évaluation exactes. Par exemple, des contraintes comme le fait d'exiger que toutes les PIDO soient positives, de permettre ou de prévenir la sélection conjointe de certaines unités, ou encore de sélectionner des unités selon un flux (par exemple ordre chronologique ou séquentiel) peuvent accroître substantiellement la complexité du plan ou de sa mise en œuvre.

Dans l'échantillonnage de population finie, l'amélioration de l'efficacité des plans d'échantillonnage exige souvent l'introduction de telles contraintes, ce qui peut compliquer les calculs des probabilités. À titre d'exemple, prenons l'échantillonnage avec probabilité proportionnelle à la taille (PPT). Dans ce type de plan avec probabilités inégales, les unités d'échantillonnage sont sélectionnées proportionnellement à la taille d'une variable auxiliaire, ce qui conduit à un plan efficace lorsque cette variable auxiliaire est corrélée avec la variable d'intérêt principale. Le plan avec probabilités inégales avec remise a été introduit pour la première fois par Hansen et Hurwitz (1943), tandis que Madow (1949), Narain (1951) et Horvitz et Thompson (1952) ont proposé des versions sans remise de l'échantillonnage avec PPT. La sélection d'un échantillon avec PPT sans remise de taille fixe peut nécessiter des procédures complexes, particulièrement lorsque des contraintes supplémentaires au niveau du plan sont imposées. À cette fin, de nombreux plans différents ont été proposés, dont 50 énumérés dans Brewer et Hanif (1983) et Tillé (2006, 2020). Tous ces plans peuvent mettre en œuvre un plan de sondage avec PPT satisfaisant les PIPO (ce qui donne une estimation sans biais de certains paramètres à l'aide de l'estimateur NHT), mais ils peuvent conduire à des PIDO différentes, ce qui a une incidence sur la précision de l'estimateur NHT.

Malgré les diverses méthodes existantes pour mettre en œuvre des plans avec probabilités inégales sans remise de taille fixe, la détermination d'un plan de sondage offrant une précision accrue sous des contraintes particulières demeure un domaine de recherche difficile et dynamique. Par exemple, la méthode du cube proposée par Deville et Tillé (2004) fournit un algorithme élégant et efficace pour l'échantillonnage équilibré, particulièrement lorsque l'équilibrage sur des variables auxiliaires est prioritaire. Plus récemment, les plans de sondage déterminantaux, étudiés par Loonis et Mary (2019) et Loonis (2023), offrent un cadre puissant et souple permettant un contrôle exact des probabilités d'inclusion grâce à la paramétrisation des matrices hermitiennes contractantes. Ces plans possèdent des structures de corrélation souhaitables inhérentes aux processus déterminantaux, lesquelles conduisent souvent à des propriétés de variance favorables des estimateurs NHT pour plusieurs variables auxiliaires. Ainsi, l'étude des combinaisons de PIDO qui produisent un échantillonnage avec PPT optimal continue de compléter ces méthodologies avancées et demeure un domaine d'intérêt majeur.

S'appuyant sur ces méthodes existantes et leurs limites, le présent article introduit une nouvelle approche visant à offrir une souplesse accrue et des capacités de conception plus intuitives. Le concept central de cette nouvelle approche consiste à proposer une méthode graphique souple pour la mise en œuvre de plans de sondage avec probabilités égales et inégales sans remise de taille fixe. Le cadre graphique proposé permet de préserver des PIPO données tout en offrant une flexibilité en ce qui concerne l'ajustement des PIDO. Alors qu'une version simplifiée de la méthode correspond à l'approche d'échantillonnage systématique de Madow (1949), les sections suivantes montrent qu'à partir d'une configuration simple (par exemple l'échantillonnage systématique de Madow) et par l'ajustement itératif de la PIDO selon une PIPO fixe, la procédure peut permettre de générer une vaste famille (habituellement continue) de plans réalisables avec des probabilités conjointes ajustées avec précision. En choisissant des règles de mise à jour ou des fonctions objectives appropriées, il est possible d'obtenir des plans de taille fixe ou d'orienter la construction vers des solutions optimisant un critère donné. Cette polyvalence permet des transitions en douceur entre des plans distincts au sein d'un cadre unique et établit un lien entre l'échantillonnage de population finie et des recherches algorithmiques intelligentes de plans optimaux (par exemple variance plus faible, meilleure dispersion spatiale ou autres critères dictés par l'application).

Voici la structure de la suite de l'article. La section 2 présente la nouvelle approche. À la section 3, on élabore des procédures permettant de générer de nouveaux plans en ajustant les PIDO selon des PIPO fixes. À la section 4, on propose une mise en œuvre à taille fixe, on démontre qu'à la base son approche est équivalente à l'approche de Madow (1949) et l'on montre la façon dont le même mécanisme d'ajustement des PIDO permet de générer d'autres plans de taille fixe. La section 5 présente une méthode intelligente de recherche d'un plan optimal. Des simulations sont effectuées à la section 6. Enfin, la section 7 conclut l'article par un résumé et des suggestions.

## 2. Approche graphique de l'échantillonnage de population finie

Dans la présente étude, une approche graphique est proposée pour l'échantillonnage avec PIPO prédéterminées, tel que présenté dans l'algorithme 1.

**Algorithme 1** Échantillonnage graphique de population finie

- 
- 1 : Créer un système de coordonnées bidimensionnel, placer les unités de population sur l'axe horizontal et considérer l'axe vertical pour les probabilités d'inclusion de premier ordre (PIPO) comprises entre 0 et 1,
  - 2 : **pour**  $k = 1, 2, \dots, N$  **faire**
  - 3 : Tracer une barre de longueur  $\pi_k$  au-dessus du point  $k$  à une position arbitraire ou aléatoire, de sorte que la barre soit entièrement comprise entre 0 et 1,
  - 4 : **fin pour**
  - 5 : Sélectionner un point aléatoire  $r$  entre 0 et 1, le reporter sur l'axe vertical, puis tracer une ligne horizontale, disons une droite aléatoire, à partir de ce point.
  - 6 : Les unités d'échantillonnage finales sont toutes celles pour lesquelles la droite aléatoire traverse les barres correspondant à leurs PIPO.
- 

La figure 2.1 présente un exemple de l'algorithme 1 avec

$$\boldsymbol{\pi} = \{0,38; 0,30; 0,42; 0,65; 0,25; 0,10; 0,90\}, \quad (2.1)$$

et chacune est représentée par une couleur différente.

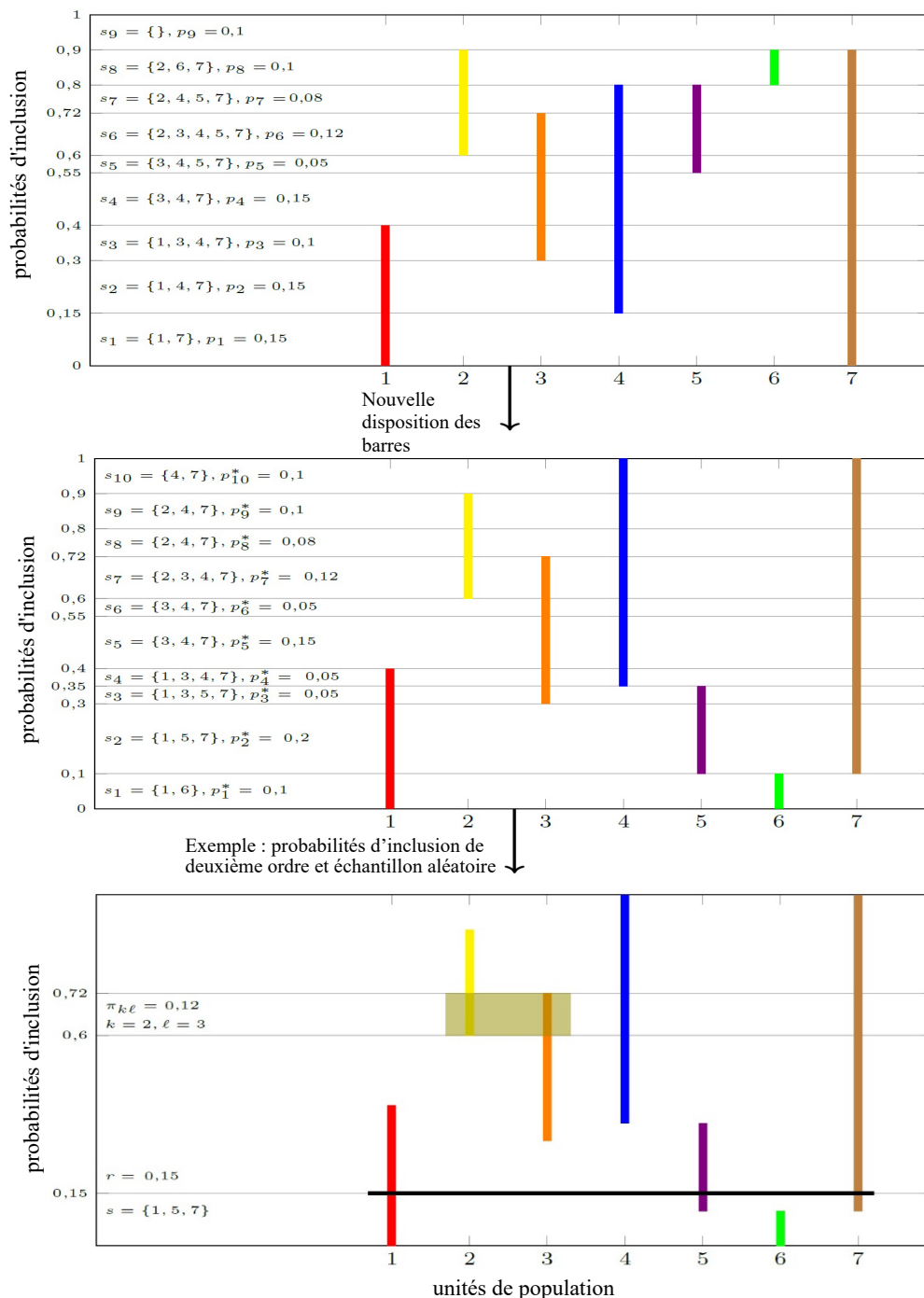
Les deux graphiques du haut de la figure 2.1 illustrent deux dispositions différentes des barres, chacune correspondant à un plan de sondage distinct. Ces deux dispositions respectent la contrainte des PIPO. Les plans s'y rattachant (échantillons et probabilités), à gauche, montrent que toute modification de la position des barres se traduit par un nouveau plan de sondage, comme l'illustre le passage du plan  $\mathbf{p}$  avec  $T = 9$  dans le graphique du haut au plan  $\mathbf{p}^*$  avec  $T = 10$  dans le graphique du milieu, où les barres pour  $\pi_4$ ,  $\pi_5$ ,  $\pi_6$  et  $\pi_7$  ont été redispesées.

Il convient de mentionner que lorsque les barres sont redispesées, la configuration initiale est traitée comme plan de référence. Pour chaque nouvelle disposition, nous ne recalculons pas tous les échantillons depuis le début; nous conservons plutôt les ensembles d'échantillons déjà générés et n'ajoutons que ceux correspondant à la nouvelle structure de barres. Cette approche permet d'éviter des coûts de calcul inutiles. Toutefois, dans certains cas, différentes dispositions des barres peuvent générer des ensembles d'échantillons identiques; il est alors possible de les fusionner, ce qui réduit la valeur effective de  $T$ . Par exemple, dans le graphique du milieu,  $s_5$  pourrait être fusionné avec  $s_6$ , et  $s_8$  pourrait l'être avec  $s_9$ , ce qui donnerait  $T = 8$ , soit essentiellement le même plan que celui représenté avec  $T = 10$ .

Le graphique du bas de la figure 2.1 présente deux exemples du plan  $\mathbf{p}^*$  pour illustrer différents concepts liés à l'échantillonnage graphique de population finie (EGPF).

- Le premier exemple montre la façon de calculer la PIDO des unités, qui peut être obtenue en déterminant l'intersection des barres. Dans cette illustration, la PIDO des unités  $k = 2$  et  $\ell = 3$  est mise en évidence en gris, ce qui donne une valeur de  $\pi_{k\ell} = 0,12$ , comme le précise le plan.
- Le deuxième exemple consiste à tracer une ligne aléatoire, représentée par une droite horizontale noire à  $r = 0,15$ . Cette droite aléatoire traverse la PIPO des unités 1, 5 et 7, ainsi ces trois unités sont sélectionnées pour faire partie de l'échantillon final. Cette représentation visuelle illustre la manière dont un processus de sélection aléatoire peut permettre de déterminer l'échantillon final en fonction de l'EGPF.

**Figure 2.1** Les deux graphiques du haut représentent différentes dispositions arbitraires de barres pour lesquelles  $\pi = \{0,38; 0,30; 0,42; 0,65; 0,25; 0,10; 0,90\}$  qui sont représentées par des couleurs différentes selon l'algorithme 1, et ils montrent que les deux dispositions respectent les probabilités d'inclusion de premier ordre spécifiées. Les plans ( $p$  et  $p^*$ ) sont calculés et présentés à gauche des graphiques, et illustrent que toute modification de la position des barres entraîne la création d'un nouveau plan. Le graphique du bas présente deux exemples : 1) le calcul des probabilités d'inclusion de deuxième ordre pour les unités  $k = 2$  et  $\ell = 3$ , surlignées en gris, ce qui donne  $\pi_{k\ell} = 0,12$ ; 2) une ligne aléatoire (droite horizontale noire) est tracée et permet de sélectionner les unités 1, 5 et 7 pour qu'elles fassent partie de l'échantillon final  $s = \{1, 5, 7\}$



À l'aide de l'EGPF, on peut obtenir un estimateur sans biais de  $Y$  tout en préservant les PIPO. Par la suite, redimensionner les barres permet de modifier les PIDO, ce qui permet de contrôler la variance de l'estimateur (et donc sa précision).

Le résultat suivant montre que la méthode préserve les PIPO.

**Théorème 2.1.** *Dans l'approche de l'EGPF, quelle que soit la disposition des barres,*

$$Pr(k \in \mathbb{S}) = \pi_k; k = 1, 2, \dots, N, \quad E(n_{\mathbb{S}}) = \sum_{k \in U} \pi_k.$$

*Démonstration.* La démonstration découle directement de l'algorithme 1.

L'algorithme 1 présente deux limitations importantes : une forte propension aux PIDO nulles et à une taille d'échantillon variable. À titre d'exemple, prenons le graphique du haut de la figure 2.1, où la taille de l'échantillon fluctue entre  $n_s = 0, 2, 3, 4, 5$  et les PIDO pour  $k=1, \ell = 2, 5, 6$  sont toutes nulles. Ces lacunes justifient un examen plus approfondi, puisqu'elles ont une incidence importante sur l'efficacité et l'applicabilité de l'algorithme dans des contextes réels. Les sections suivantes présentent de nouveaux algorithmes conçus expressément pour surmonter ces limitations et améliorer de façon considérable l'efficacité globale du processus d'échantillonnage.

### 3. Génération de nouveaux plans par l'ajustement des probabilités d'inclusion de deuxième ordre

Même un seul élément nul dans la matrice de PIDO empêche l'existence d'un estimateur de la variance sans biais pour l'estimateur NHT. Par conséquent, même si un plan ayant une matrice de PIDO strictement positive ne garantit pas une efficacité supérieure, il est toujours intéressant de construire de tels plans, car ils permettent une estimation de la variance sans biais. Dans la présente section, un algorithme est proposé pour la réduction du nombre d'éléments nuls dans la matrice de PIDO.

Pour commencer, il convient de mentionner que chaque barre représente initialement la portion totale associée à une  $\pi_k$  donnée. Une barre peut être divisée en petites parties appelées segments. Ces segments indiquent les différents échantillons auxquels une unité peut appartenir (voir la figure 3.1). Considérons ensuite l'axe des  $y$  partitionné en  $D$  bandes, désignées par  $\Delta_1, \Delta_2, \dots, \Delta_D$ , correspondant aux échantillons  $s_1, s_2, \dots, s_D$ . Il faut souligner que  $D \geq T$ , puisque certains échantillons peuvent apparaître plusieurs fois tout au long du partitionnement de l'axe des  $y$ . Sélectionnons maintenant deux bandes,  $\Delta_i$  et  $\Delta_j$ , de hauteur respective  $p_i$  et  $p_j$ . À l'intérieur de chaque bande, considérons deux sous-bandes plus petites  $\delta_i$  et  $\delta_j$ , toutes deux de taille  $v_{i,j}(\alpha) = \alpha \times \min(p_i, p_j)$  pour  $\alpha \in [0, 1]$ , et définissons :

$$\delta_i(k) = \begin{cases} 1 & \text{Si la barre pour l'unité } k \text{ couvre entièrement la hauteur de la sous-bande } \delta_i, \\ 0 & \text{autrement.} \end{cases}$$

Après, si  $s_{i-j} = s_i \setminus s_j$  n'est pas vide, sélectionnons  $k \in s_{i-j}$  et remplaçons le segment de  $\pi_k$  dans  $\delta_i$  par  $\delta_j$ . Autrement dit, posons  $\delta_i(k) = 0$  et  $\delta_j(k) = 1$ , ce qui peut mener aux résultats suivants :

- l'échantillon  $s_i$  de probabilité  $p_i$  se scinde en deux échantillons :

$$1 : s_{i,1} = s_i, \quad p_{i,1} = p_i - v_{i,j}(\alpha),$$

$$2 : s_{i,2} = s_i \setminus \{k\}, \quad p_{i,2} = v_{i,j}(\alpha),$$

- l'échantillon  $s_j$  de probabilité  $p_j$  se scinde en deux échantillons :

$$3 : s_{j,1} = s_j, \quad p_{j,1} = p_j - v_{i,j}(\alpha),$$

$$4 : s_{j,2} = s_j \cup \{k\}, \quad p_{j,2} = v_{i,j}(\alpha).$$

En résumé, pour potentiellement réduire les PIDO nulles et générer de nouveaux plans, l'algorithme 1 consiste à sélectionner aléatoirement un segment d'une barre et à le déplacer vers une nouvelle position, à condition qu'un emplacement vacant soit disponible. Des précisions sont fournies dans l'algorithme 2.

Avant de présenter cet algorithme, il est important de préciser l'utilisation du terme « aléatoirement » dans ce contexte. La sélection aléatoire de deux bandes fait référence à un mécanisme aléatoire pouvant être mis en œuvre de différentes façons. Dans les simulations figurant dans le présent article, on utilise un échantillon aléatoire simple de taille deux parmi toutes les paires de bandes possibles. La sélection peut également être effectuée avec des probabilités proportionnelles à la taille, où la valeur  $p_i$  désigne la hauteur de la bande  $i$ , laquelle représente à son tour le poids de sondage de l'échantillon correspondant. Le choix de la méthode de sélection peut influencer le taux de convergence de l'algorithme vers les propriétés visées du plan.

---

**Algorithme 2** Échantillonnage graphique de population finie chaotique : générateur de plans

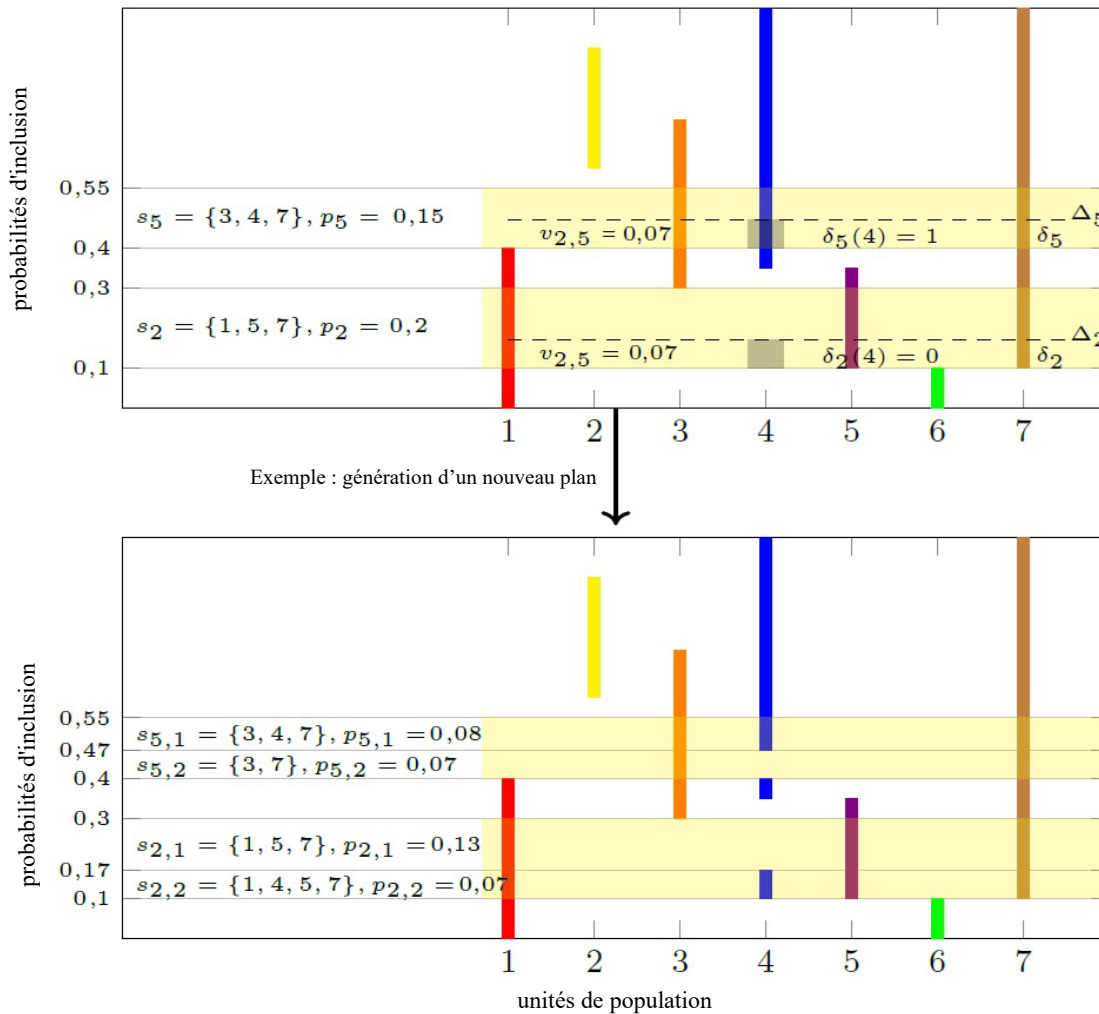
---

- 1 : Créer un système de coordonnées bidimensionnel, en plaçant les unités de population sur l'axe horizontal et en utilisant l'axe vertical pour les probabilités d'inclusion de premier ordre (PIPO), allant de 0 à 1, puis construire le plan de sondage.
  - 2 : Choisir le nombre total d'itérations chaotiques  $M$ .
  - 3 : **pour**  $m = 1, 2, \dots, M$  **faire**
  - 4 : Choisir aléatoirement deux bandes  $\Delta_i$  et  $\Delta_j$ , et à l'intérieur de celles-ci, deux sous-bandes  $\delta_i$  et  $\delta_j$  de taille  $v_{i,j}(\alpha)$  avec  $0 < \alpha < 1$ .
  - 5 : **si**  $s_{i-j}$  n'est pas vide **alors**
  - 6 : Sélectionner aléatoirement  $k \in s_{i-j}$ .
  - 7 : Fixer  $\delta_i(k) = 0$  et  $\delta_j(k) = 1$ .
  - 8 : **fin si**
  - 9 : **fin pour**
  - 10 : Tracer une droite aléatoire.
  - 11 : L'échantillon final comprend toutes les unités pour lesquelles la droite aléatoire traverse leur barre respective associée aux PIPO.
- 

Pour expliquer le concept de base de l'algorithme 2, examinons les graphiques de la figure 3.1. Dans le graphique du haut, considérons  $\Delta_2$  et  $\Delta_5$  et observons qu'à l'intérieur de celles-ci,  $\delta_5(4) = 1$  et  $\delta_2(4) = 0$ . On peut alors tronquer le segment correspondant de  $\pi_4$  dans  $\delta_5$  et le déplacer vers la nouvelle position dans  $\delta_2$ , ce qui donne  $\delta_2(4) = 1$  et  $\delta_5(4) = 0$ , comme l'illustre le graphique du bas. Cette mise en œuvre de l'algorithme 2 avec  $M = 1$  permet de générer un nouveau plan comportant un plus grand nombre d'échantillons possibles. Par exemple, dans le graphique du haut,  $\pi_{k\ell} = 0$  pour  $k = 4$  et  $\ell = 5$ , alors que dans le nouveau plan, on obtient  $\pi_{k\ell} = 0,07$  pour  $k = 4$  et  $\ell = 5$ .

Il convient de mentionner que, dans l'EGPF, les PIDO sont exactement calculables. Pour tout niveau de raffinement  $M$ , l'algorithme 2 s'exécute en maintenant les PIPO fixes jusqu'à ce que la configuration se stabilise; l'échantillonnage est ensuite effectué à partir de ce plan final, et les PIDO opérationnelles sont celles induites par cette configuration stabilisée.

**Figure 3.1** Exemple de génération d'un nouveau plan de sondage  $p^*$  (graphique du milieu de la figure 2.1). En sélectionnant  $\Delta_2$  et  $\Delta_5$  de hauteur  $p_2^* = 0,2$  et  $p_5^* = 0,15$ , et en choisissant des sous-bandes  $\delta_2$  et  $\delta_5$  de taille  $v_{2,5}(7/15) = 0,07$ , il est possible d'échanger des segments entre les sous-bandes  $\delta_2$  et  $\delta_5$  pour l'unité  $k = 4$ , ce qui donne 4 échantillons :  $s_{2,1}$ ,  $s_{2,2}$ ,  $s_{5,1}$  et  $s_{5,2}$



Dans l'algorithme 2, chaque échange de segments préservant les PIPO disperse de façon empirique la masse commune sur un plus grand ensemble d'échantillons réalisables et tend à réduire le nombre d'entrées nulles dans la matrice de PIDO. Bien que l'on ne revendique pas d'amélioration monotone à chaque étape, ce comportement semble indiquer une augmentation progressive de la propagation distributionnelle de la distribution, ce qui motive la conjecture suivante pour des travaux futurs portant sur cette classe d'algorithmes.

**Conjecture 3.1.** *L'EGPF chaotique converge vers l'échantillonnage de Poisson lorsque  $M \rightarrow \infty$ .*

## 4. Algorithme à taille fixe d'échantillonnage graphique de population finie

À partir de la résolution du problème de PIDO nulles présentée à la section précédente, la présente section aborde le défi posé par la taille d'échantillon aléatoire dans le cadre de l'algorithme 1.

Pour relever ce défi, l'une des approches les plus simples est de placer les barres de façon cumulative comme suit. La première barre s'étend de zéro à  $\pi_1$ , la deuxième barre s'étend de  $\pi_1$  à  $\min(\pi_1 + \pi_2, 1)$ , et si le minimum est de 1, le reste de  $(\pi_1 + \pi_2) - 1$  repart à zéro, et ainsi de suite. Des précisions sont présentées dans l'algorithme 3.

---

### Algorithme 3 Échantillonnage graphique de population finie à taille fixe, équivalent à Madow (1949)

---

```

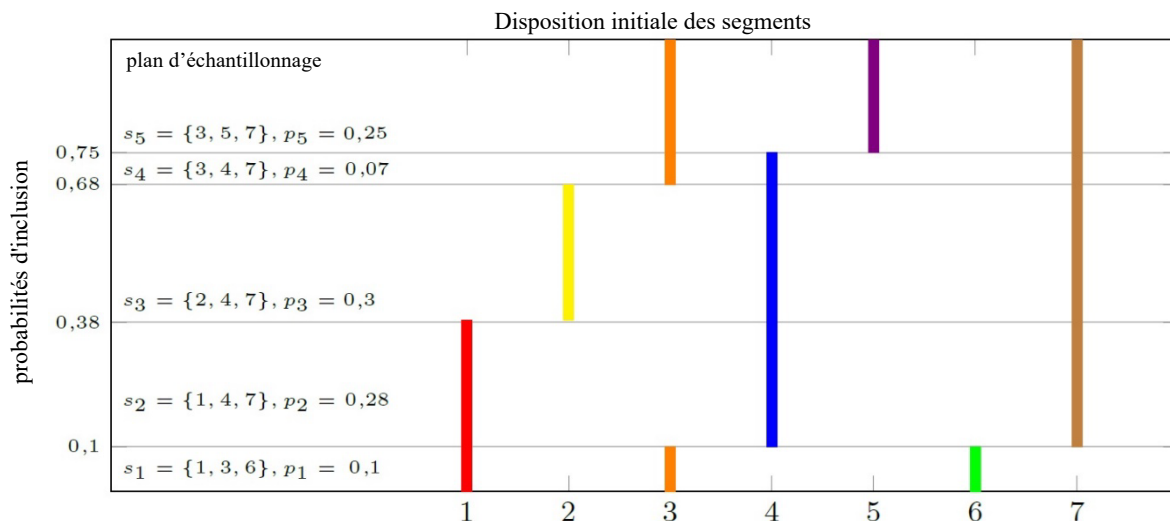
1 : La première barre couvre  $[0, \pi_1]$ , fixer  $b_1 = \pi_1$ ,
2 : pour  $k = 2, 3, \dots, N$  faire
3 :   si  $\pi_k + b_{k-1} \leq 1$ , alors la  $k^{\text{e}}$  barre couvre  $(b_{k-1}, \pi_k + b_{k-1}]$ , fixer  $b_k = \pi_k + b_{k-1}$ ,
4 :   sinon la  $k^{\text{e}}$  barre couvre  $(b_{k-1}, 1] \cup [0, \pi_k + b_{k-1} - 1]$ , fixer  $b_k = \pi_k + b_{k-1} - 1$ ,
5 :   fin si
6 : fin pour

```

---

La figure 4.1 présente une version à taille fixe de l'EGPF avec  $\pi = \{0,38; 0,30; 0,42; 0,65; 0,25; 0,10; 0,90\}$  disposées séquentiellement pour former un plan de taille fixe  $n = \sum_{k=1}^7 \pi_k = 3$ . Dans cette disposition, les PIPO sont harmonisées pour maintenir la contrainte de taille fixe.

**Figure 4.1** Le graphique présente une version à taille fixe (algorithme 3) de l'échantillonnage graphique de population finie pour laquelle  $\pi = \{0,38; 0,30; 0,42; 0,65; 0,25; 0,10; 0,90\}$  qui sont disposées séquentiellement pour créer un plan de taille fixe



Fait intéressant, l'algorithme 3 peut être interprété comme une adaptation verticale de la méthode d'échantillonnage systématique proposée par Madow (1949). Toutefois, comme il sera démontré ci-après, contrairement à l'échantillonnage systématique décrit par Madow (1949), cette approche novatrice permet

de reproduire divers plans d'échantillonnage traditionnels et nouveaux, y compris de nombreuses méthodes d'échantillonnage de taille fixe avec probabilités inégales.

Pour générer un nouveau plan à partir de la version de l'EGPF de Madow tout en préservant les PIPO et la contrainte de taille fixe, une stratégie semblable à celle décrite à la section 3 peut être utilisée. Plus précisément, deux bandes correspondant à deux échantillons peuvent être sélectionnées, et deux segments de leurs barres peuvent être intervertis, à condition qu'un espace vide existe aux nouvelles positions. Cette approche permet de préserver les tailles d'échantillon fixes tout en générant un nouveau plan. Toutefois, pour présenter un algorithme complet pour ce processus, il est nécessaire de définir deux concepts clés.

*Définition 4.1.* Considérons deux sous-bandes sur l'axe vertical,  $\delta_i \subseteq \Delta_i$  et  $\delta_j \subseteq \Delta_j$ . Deux segments,  $\delta_i(k)$  et  $\delta_j(\ell)$ , sont interchangeables si

$$\delta_i(k) = 1, \quad \delta_j(\ell) = 1; \quad \delta_j(k) = 0, \quad \delta_i(\ell) = 0.$$

*Définition 4.2.* Intervertir deux segments interchangeables  $\delta_i(k)$  et  $\delta_j(\ell)$  consiste à poser

$$\delta_i(k) = 0, \quad \delta_j(\ell) = 0; \quad \delta_j(k) = 1, \quad \delta_i(\ell) = 1.$$

Considérons maintenant deux bandes  $\Delta_i$  et  $\Delta_j$  de hauteur  $p_i$  et  $p_j$ , et deux sous-bandes  $\delta_i$  et  $\delta_j$  de même taille  $v_{i,j}(\alpha)$ . Si  $s_{i-j}$  et  $s_{j-i}$  ne sont pas vides, on sélectionne  $k \in s_{i-j}$  et  $\ell \in s_{j-i}$  aléatoirement et on les intervertit, ce qui peut mener à :

- l'échantillon  $s_i$  de probabilité  $p_i$  se scinde en deux échantillons :

$$\begin{aligned} 1 : \quad s_{i,1} &= s_i, & p_{i,1} &= p_i - v_{i,j}(\alpha), \\ 2 : \quad s_{i,2} &= s_i \cup \{\ell\} \setminus \{k\}, & p_{i,2} &= v_{i,j}(\alpha), \end{aligned}$$

- l'échantillon  $s_j$  de probabilité  $p_j$  se scinde en deux échantillons :

$$\begin{aligned} 3 : \quad s_{j,1} &= s_j, & p_{j,1} &= p_j - v_{i,j}(\alpha), \\ 4 : \quad s_{j,2} &= s_j \cup \{k\} \setminus \{\ell\}, & p_{j,2} &= v_{i,j}(\alpha). \end{aligned}$$

Des précisions sont présentées dans l'algorithme 4.

---

**Algorithme 4** Échantillonnage graphique de population finie chaotique à taille fixe

---

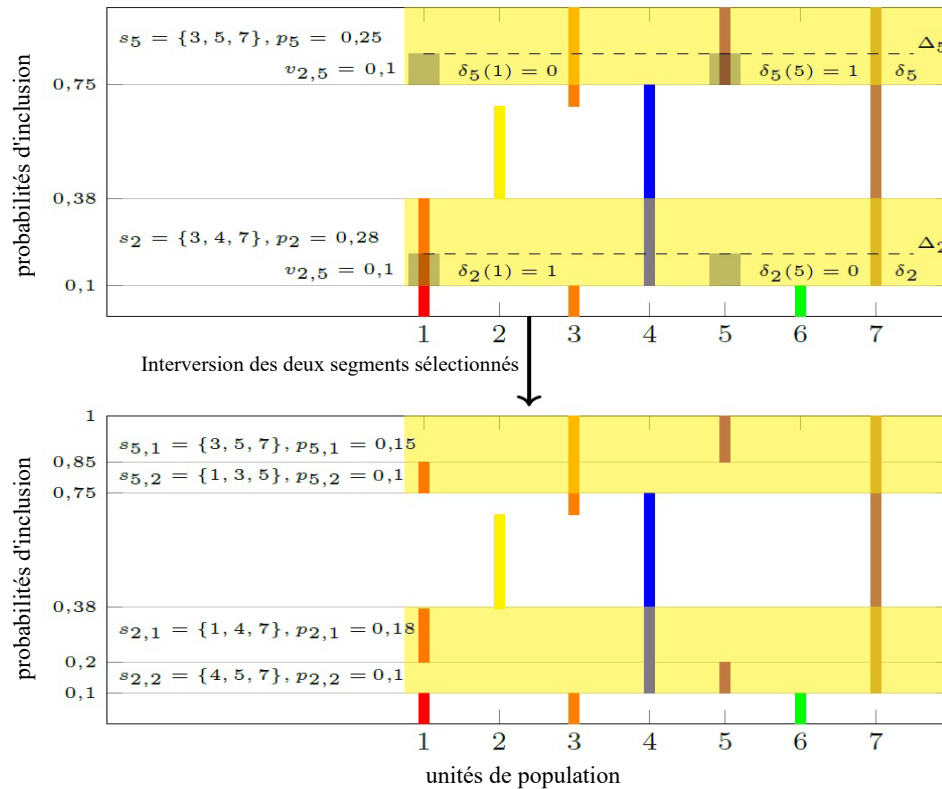
```

1 : Mettre en œuvre l'algorithme 3.
2 : pour  $m = 1, 2, \dots, M$  faire
3 :   Sélectionner aléatoirement deux bandes  $\Delta_i$  et  $\Delta_j$ , et, à l'intérieur de celles-ci, deux sous-bandes  $\delta_i$  et  $\delta_j$  de taille  $v_{i,j}(\alpha)$ ,  $0 < \alpha < 1$ ,
4 :   si  $s_{i-j}$  et  $s_{j-i}$  ne sont pas vides alors
5 :     sélectionner  $k \in s_{i-j}$  et  $\ell \in s_{j-i}$  aléatoirement,
6 :     intervertir  $\delta_i(k)$  et  $\delta_j(\ell)$ .
7 :   fin si
8 : fin pour
```

---

À titre d'exemple de l'algorithme 4, considérons la figure 4.2, dans laquelle deux bandes,  $\Delta_2$  et  $\Delta_5$ , sont sélectionnées et, à l'aide de deux sous-bandes de taille  $v_{2,5}(10/25) = 0,1$ , des segments des unités 1 et 5 sont intervertis pour générer un nouveau plan.

**Figure 4.2** Le graphique du haut présente une version à taille fixe de l'échantillonnage graphique de population finie avec  $\pi = \{0,38; 0,30; 0,42; 0,65; 0,25; 0,10; 0,90\}$  disposées séquentiellement pour créer un plan de taille fixe dans lequel deux bandes,  $\Delta_2$  et  $\Delta_5$ , sont sélectionnées. À l'intérieur de ces bandes, deux sous-bandes de taille  $v_{2,5} = v_{2,5}(10/25) = 0,10$  sont indiquées par des lignes pointillées. Les unités 1 et 5 étant interchangeables dans cette situation, elles sont interverties, ce qui donne deux nouveaux échantillons,  $s_{2,2}$  et  $s_{5,2}$ , en plus des échantillons originaux,  $s_{2,1}$  et  $s_{5,1}$ , représentés dans le graphique du bas



Le théorème suivant énonce les invariants de base de la procédure chaotique de taille fixe pour l'algorithme 4.

**Théorème 4.3.** Selon l'approche d'EGPF chaotique à taille fixe fondée sur l'algorithme 4,

$$Pr(k \in \mathbb{S}) = \pi_k; \quad k=1, 2, \dots, N,$$

$$E(n_{\mathbb{S}}) = \sum_{k \in U} \pi_k,$$

et si  $E(n_{\mathbb{S}})$  est un entier, le plan est à taille fixe; de manière équivalente :

$$var(n_{\mathbb{S}}) = 0.$$

*Démonstration.* La démonstration découle directement de l'algorithme 4.

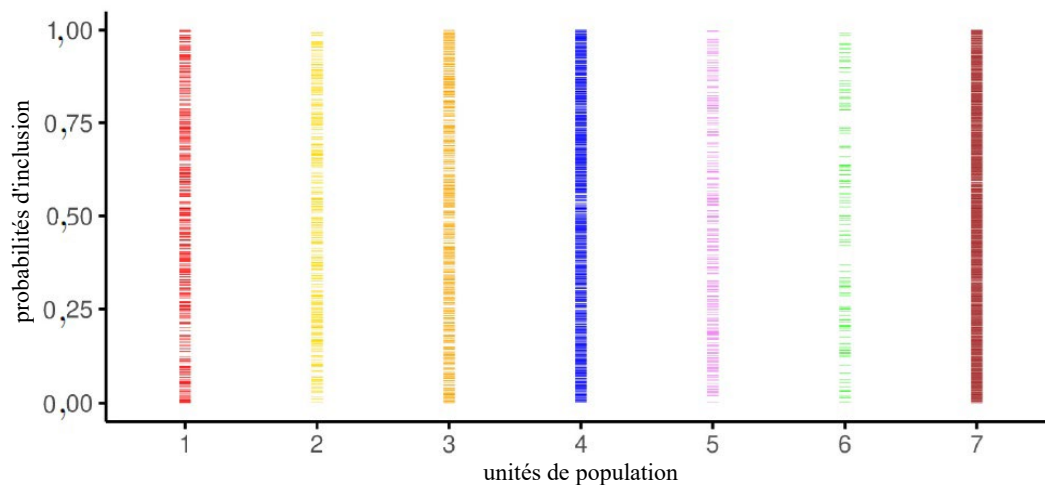
Parallèlement à la conjecture 3.1 et en notant que la procédure chaotique à taille fixe permet de préserver à la fois les PIPO prescrites et la taille d'échantillon tout en répartissant de façon empirique la masse de probabilité sur l'ensemble des échantillons possibles de taille  $n$ , il semble qu'elle accroisse l'entropie

soumise à la contrainte  $|\mathcal{S}| = n$ . Cela motive la conjecture suivante selon laquelle le plan limitatif coïncide avec le plan de Poisson conditionnel (l'analogue à taille fixe de l'échantillonnage de Poisson, couramment décrit comme un échantillonnage à entropie maximale (Tillé, 2006)).

**Conjecture 4.4.** *L'EGPF chaotique à taille fixe avec  $M \rightarrow \infty$  converge vers un échantillonnage à entropie maximale.*

La figure 4.3 illustre un plan de taille fixe correspondant au vecteur de PIPO (2.1), généré par l'algorithme 4 avec  $\alpha = 0,5$  et  $M = 10\,000$ , lequel est caractérisé par des probabilités de plan largement distribuées et par un nombre nettement plus faible d'entrées de PIDO nulles.

**Figure 4.3** Mise en œuvre de l'algorithme 4 sur les données de population (2.1) avec  $\alpha = 0,5$  et  $M = 10\,000$



Le cadre proposé préserve les PIPO prescrites, appuie un échantillonnage à taille fixe et permet l'ajustement des PIDO au moyen de mises à jour locales, ce qui génère ainsi une vaste famille de plans réalisables dont les PIDO sont exactement calculables. Ces propriétés font qu'il est naturel d'intégrer la construction au sein d'une boucle d'optimisation cherchant des plans harmonisés à des objectifs liés à des problèmes particuliers, tout en respectant les contraintes prescrites.

## 5. Échantillonnage graphique de population finie optimal

Comme l'ont souligné Thompson et Seber (1996), il n'existe pas de plan d'échantillonnage globalement optimal pour l'échantillonnage de population finie. Le caractère optimal d'un plan dépend de plusieurs facteurs, notamment de la structure de la variable principale, des PIPO et de l'estimateur utilisé. Lorsqu'il

est question de plan optimal, l'accent est mis sur la façon dont, étant donné un vecteur de PIPO, il est possible de mettre en œuvre une méthode d'échantillonnage avec probabilités inégales dépassant les plans d'échantillonnage conventionnels bien connus. Il convient également de mentionner que le présent examen vise principalement à améliorer l'étape de conception lorsque l'estimateur NHT est utilisé. De futurs travaux pourraient porter sur l'amélioration à la fois des étapes de conception et d'estimation à l'aide de l'EGPF, ce qui constituerait une piste de recherche intéressante.

Avec l'EGPF optimal, la capacité de l'EGPF à générer divers plans illustre une nouvelle voie susceptible d'offrir un moyen simple et efficace de découvrir de nouveaux plans ayant un bon rendement.

Dans l'EGPF chaotique (à taille aléatoire ou fixe), les PIPO demeurent fixes, ce qui facilite la création de plans comportant un ensemble très vaste et souple de PIDO possibles par une nouvelle disposition des barres. En tirant parti des variables auxiliaires appropriées, il devient possible de construire un grand nombre de plans. Une fois la position des barres (segments) fixée, le plan est entièrement déterminé dans le cadre de l'EGPF, ce qui permet de calculer des critères tels que la variance, l'erreur maximale, etc., de l'estimateur NHT en fonction de certaines variables auxiliaires. Par conséquent, une approche consiste à fixer un critère, à redimensionner les barres pour modifier les PIDO, à générer les plans candidats, puis à sélectionner le meilleur selon ce critère.

Dans le cas le plus simple, considérons deux variables auxiliaires  $x$  et  $z$  présentant des corrélations raisonnables avec  $y$ ; l'une sert à construire les PIPO et l'autre, à évaluer les plans en fonction de certains critères. L'utilisation de

$$\pi_k = \min (cx_k, 1), \quad \text{où} \quad \sum_{k \in U} \pi_k = n, \quad (5.1)$$

conduit à des PIPO fondées sur  $x$  de taille d'échantillon  $n$ , ce qui correspond à des PPT. Dans ce plan, on sait que  $\text{var}(\hat{X}) = 0$ . En utilisant ensuite une variable d'évaluation  $z$  et en considérant un paramètre, par exemple  $\theta_z$ , le fait de redimensionner les barres un grand nombre de fois à l'aide d'un algorithme aléatoire ou intelligent, puis de sélectionner la meilleure disposition pour estimer  $\theta_z$  à l'aide de l'estimateur NHT de  $z$ , permet d'aboutir à la sélection d'un plan optimal.

Tout algorithme de recherche parmi des plans d'échantillonnage nécessite un critère permettant d'évaluer sa proximité au résultat souhaité. Dans le contexte de l'optimisation de l'EGPF, divers paramètres du plan peuvent être retenus comme critères. Les équations suivantes présentent trois critères possibles pour l'EGPF optimal :

$$C_1 (\theta_z = Z, \mathbf{p}) = \sum_{t=1}^T (\hat{Z}_{s_t} - Z)^2 p_t,$$

$$C_2 (\theta_z = Z, \mathbf{p}) = \sum_{t=1}^T |\hat{Z}_{s_t} - Z| p_t,$$

$$C_3 (\theta_z = Z, \mathbf{p}) = \max_{t=1, \dots, T} |\hat{Z}_{s_t} - Z|,$$

où

$$Z = \sum_{k=1}^N z_k, \quad \hat{Z}_{s_t} = \sum_{k \in s_t} \frac{z_k}{\pi_k}.$$

Il convient de mentionner que malgré les similitudes conceptuelles du processus d'optimisation de l'EGPF à l'aide d'une variable auxiliaire  $z$  comme cible d'équilibrage avec les méthodes d'échantillonnage équilibré (Deville et Tillé, 2004; Loonis, 2023), l'EGPF optimal n'impose pas explicitement de contraintes d'équilibre à chaque étape. Il vise plutôt à cerner des plans réduisant au minimum un critère donné (comme la variance ou l'erreur maximale) au moyen d'une nouvelle disposition intelligente des segments, ce qui peut mener à des échantillons équilibrés ou quasi équilibrés comme résultat indirect, plutôt que comme exigence explicite.

### ***Algorithme de recherche de plan de sondage intelligent***

Une variante d'une approche par l'algorithme de recherche glouton du meilleur à prime abord est présentée pour trouver un plan optimal. Elle commence par l'initialisation de variables auxiliaires :  $x$  pour construire les PIPO et  $z$  pour évaluer les plans. Un plan initial est ensuite établi en fonction des PIPO, servant de point de départ au processus d'optimisation.

L'algorithme effectue plusieurs itérations, disons  $\Lambda$ , au cours desquelles, à chaque cycle, un nombre  $N_{\text{nœud}}$  de nouveaux plans candidats sont générés en modifiant le dernier plan en fonction de l'algorithme 4. Cette modification permet d'examiner diverses configurations. Chaque plan nouvellement généré est évalué selon un critère prédéfini, désigné par  $C(\theta_z = Z, \mathbf{p})$ . Au départ, le plan initial est considéré comme le meilleur plan. Chaque fois qu'un plan supérieur est relevé, il remplace le meilleur plan précédent.

Deux ensembles clés sont établis au cours de ce processus : l'ensemble ouvert et l'ensemble fermé. L'ensemble ouvert contient les plans, classés par ordre de priorité selon leur coût, qui n'ont pas encore été examinés. À l'inverse, l'ensemble fermé consigne les plans déjà évalués, ce qui évite que l'algorithme traite de façon redondante un même plan plusieurs fois. Tout au long des itérations, l'algorithme suit le plan ayant le meilleur rendement. Par la suite, l'ensemble de plans potentiels est géré de façon à ne conserver que les candidats les plus prometteurs, et les plans moins efficaces sont remplacés au besoin. Cette gestion permet de mettre l'accent sur les options les plus viables en vue d'un affinage ultérieur.

Une fois toutes les itérations terminées, l'algorithme retourne le meilleur plan trouvé, représentant la configuration optimale selon le critère défini. Pour obtenir de plus amples renseignements, veuillez consulter l'algorithme 5.

---

**Algorithme 5** Échantillonnage graphique de population finie optimal : variante de l'algorithme de recherche glouton du meilleur à prime abord
 

---

```

1 : Intrants : variable auxiliaire  $x$  (pour construire les PIPO), variable d'évaluation  $z$ , critère  $C(\theta_z = Z, \mathbf{p})$  à réduire au
   minimum, budget d'itérations  $\Lambda$ , facteur de ramification  $N_{\text{nœud}}$  et capacité  $\text{max\_open\_set\_size}$ .
2 : Créer  $\text{initial\_design}$  à partir des PIPO à l'aide de l'algorithme 3.
3 : Initialiser  $\text{open\_set} \leftarrow \{\}$  (file de priorité ajustée par le critère) et  $\text{closed\_set} \leftarrow \{\}$  (plans évalués).
4 : Calculer  $\text{init\_criterion} \leftarrow C(\theta_z = Z, \mathbf{p}(\text{initial\_design}))$ .
5 : Insérer( $\text{init\_criterion}$ ,  $\text{initial\_design}$ ) dans  $\text{open\_set}$ .
6 :  $\text{best\_design} \leftarrow \text{initial\_design}$ ;  $\text{best\_criterion} \leftarrow \text{init\_criterion}$ .
7 : pour  $\text{iterations} = 1$  à  $\Lambda$  faire
8 :   Extraire et supprimer( $\text{current\_criterion}$ ,  $\text{current\_design}$ ) ayant le critère minimal de  $\text{open\_set}$ .
9 :   pour  $i = 1$  à  $N_{\text{nœud}}$  faire
10 :    Générer  $\text{new\_design} \leftarrow \text{APPLIQUER L'ALGORITHME 4}(\text{current\_design})$ 
11 :     $\text{new\_criterion} \leftarrow C(\theta_z = Z, \mathbf{p}(\text{new\_design}))$ 
12 :    si  $\text{new\_design} \notin \text{closed\_set}$  alors
13 :      si  $\text{new\_criterion} < \text{best\_criterion}$  alors
14 :         $\text{best\_design} \leftarrow \text{new\_design}$ ;  $\text{best\_criterion} \leftarrow \text{new\_criterion}$ 
15 :      fin si
16 :      si  $|\text{open\_set}| < \text{max\_open\_set\_size}$  alors
17 :        Insérer( $\text{new\_criterion}$ ,  $\text{new\_design}$ ) dans  $\text{open\_set}$ 
18 :      sinon
19 :        Soit( $\text{worst\_crit}$ ,  $\text{worst\_des}$ ), l'élément ayant le critère maximal dans  $\text{open\_set}$ 
20 :        si  $\text{new\_criterion} < \text{worst\_crit}$  alors
21 :          Remplacer( $\text{worst\_crit}$ ,  $\text{worst\_des}$ ) par( $\text{new\_criterion}$ ,  $\text{new\_design}$ ) dans  $\text{open\_set}$ 
22 :        fin si
23 :      fin si
24 :    fin si
25 :  fin pour
26 :  Ajouter  $\text{current\_design}$  à  $\text{closed\_set}$ 
27 : fin pour
28 : Retourner  $\text{best\_design}$ 

```

---

## 6. Simulations

Afin d'évaluer la capacité de l'algorithme 5 à trouver des plans d'échantillonnage optimaux, une série d'études par simulation ont été menées à partir de données synthétiques et de données réelles. Le but était de comparer l'efficacité de l'EGPF optimal par rapport aux plans suivants :

- EAS : échantillonnage aléatoire simple;
- CUB : méthode du cube (Dewille et Tillé, 2004);
- PED : plan d'échantillonnage déterminantal (Loonis, 2023).

L'EGPF optimal a été optimisé à l'aide d'un critère de variance :

$$C(\theta_z = Z, \mathbf{p}) = \sum_{t=1}^T (\hat{Z}_{s_t} - Z)^2 p_t,$$

où  $\hat{Z}_{s_t}$  est l'estimateur NHT pour  $Z$  fondé sur l'échantillon  $s_t$ .

L'efficacité des estimateurs de  $Y$  a été mesurée comme suit :

$$EF_{EAS} = \frac{\text{var}(\hat{Y}_{EAS})}{\text{var}(\hat{Y}_{OGFS})}, \quad EF_{PED} = \frac{\text{var}(\hat{Y}_{PED})}{\text{var}(\hat{Y}_{OGFS})}, \quad EF_{CUB} = \frac{\text{var}_{MC}(\hat{Y}_{CUB})}{\text{var}(\hat{Y}_{OGFS})},$$

et des expressions analogues ont été utilisées pour l'efficacité des estimateurs de  $Z$ . Les estimateurs  $\hat{Y}_{EAS}$ ,  $\hat{Y}_{PED}$  et  $\hat{Y}_{CUB}$  sont des estimateurs NHT pour les méthodes d'échantillonnage EAS, PED et CUB. Pour les méthodes EAS, PED et EGPf optimal, la variance de l'estimateur NHT peut être calculée exactement à partir des PIPO et des PIDO connues, sans avoir à recourir à une approximation de Monte Carlo. Les valeurs rapportées pour ces méthodes sont donc exactes. Pour la méthode du cube, toutefois, la variance a été approximée par simulation de Monte Carlo, car le calcul exact n'était pas simple.

Pour cette simulation, paquet spécialisé en Python, *graphical-sampling* (Panahbehagh, Mohebbi, HosseiniNasab et Moghadam, 2025), mettant en œuvre l'approche d'échantillonnage graphique, a été utilisé. Ce paquet a servi de cadre de base aux simulations de la présente section, ce qui a permis la mise en œuvre efficace et précise des techniques d'échantillonnage examinées.

Tous les plans d'échantillonnage présentés dans la présente section de simulations (à l'exception de l'EAS) reposent sur la même variable auxiliaire à la fois pour définir les probabilités d'inclusion inégales et pour assurer l'équilibrage (ou l'optimisation) du plan d'échantillonnage. Dans le cas de la méthode du cube, le vecteur de PIPO est également intégré en tant que variable d'équilibrage afin d'imposer une taille d'échantillon fixe.

Par ailleurs, dans toutes les simulations, les paramètres de l'algorithme 5 ont été fixés à  $M \sim \text{UD}(1, 3)$  et à  $\alpha \sim \text{UC}(0,7; 0,9)$ , où UD et UC désignent respectivement des distributions uniforme discrète et uniforme continue.

Ces plages de paramètres reflètent le comportement recherché de l'algorithme intelligent. Le paramètre  $M$  tient compte du nombre de modifications locales appliquées à chaque itération. Des valeurs élevées rendent la recherche instable, tandis que  $M = 1$  entraîne une exploration très lente. La randomisation de  $M$  selon UD(1, 3) permet d'avoir un équilibre pratique entre l'exploration et la stabilité. Le paramètre  $\alpha$  régit l'intensité de chaque perturbation locale dans l'algorithme 5. Des valeurs faibles produisent des progrès négligeables, tandis que des valeurs proches de 1 entraînent des mises à jour trop agressives. Le choix de  $\alpha$  dans UC(0,7; 0,9) permet d'obtenir des étapes suffisamment solides tout en demeurant stables, et ces paramètres se sont révélés robustes lors de l'ensemble des expériences préliminaires. En pratique,  $M$  devrait réguler le nombre de changements appliqués par itération, tandis que  $\alpha$  devrait en contrôler l'intensité.

## 6.1 Population simulée

Pour la population simulée, les données ont été générées comme suit :

- Taille de la population :  $N = 100$ ; taille de l'échantillon :  $n = 10$ .
- La variable principale  $y$  a été générée sous la forme  $y \sim \mathcal{N}(100, 10^2)$ .

- Les variables auxiliaires ont été générées sous la forme

$$z = b_1 y + \epsilon_z, \quad x = b_2 y + \epsilon_x, \quad b_1, b_2 \in \mathbb{R}, \quad \epsilon_z \sim \mathcal{N}(0, \sigma_z^2), \quad \epsilon_x \sim \mathcal{N}(0, \sigma_x^2),$$

où  $\sigma_z$  et  $\sigma_x$  ont été choisis de manière à obtenir des corrélations cibles approximatives de

$$\rho_{z,y}, \rho_{x,y} \in \{0,75; 0,85; 0,95\}.$$

- Pour  $x$ , un vecteur à valeur constante a également été considéré dans un scénario pour mettre en œuvre un échantillonnage avec probabilités égales lors de l'évaluation de l'EGPF optimal.

Pour chaque combinaison de  $\rho_{z,y}$  et de  $\rho_{x,y}$ , l'algorithme 5 a été exécuté avec  $\Lambda = 1\,000$  itérations et  $N_{\text{noeud}} = 30$ . Les résultats sont présentés à la figure 6.1. L'efficacité de l'EGPF optimal par rapport aux méthodes PED et CUB est illustrée au moyen de diagrammes à barres, tandis que son efficacité par rapport à la méthode EAS est indiquée sous forme de valeurs annotées au-dessus des barres. Cette présentation a été retenue parce que les valeurs d'efficacité élevées par rapport à la méthode EAS auraient donné lieu à des barres disproportionnellement hautes rendant difficile la visualisation du rendement comparatif des méthodes PED et CUB.

L'efficacité a été évaluée pour différentes combinaisons de corrélations entre les variables auxiliaires  $x$  et  $z$  et la variable principale  $y$ , où  $\rho_{x,y} = 0,00$  correspond au cas d'un échantillonnage avec probabilités égales. Bien que le but principal de l'échantillonnage soit l'estimation de  $Y$ , dans la présente étude, l'estimation de  $Z$  (qui sert de variable auxiliaire utilisée pour optimiser le plan dans l'EGPF optimal et pour imposer des contraintes d'équilibrage dans les méthodes PED et CUB) est également calculée et représentée graphiquement. Cela permet d'évaluer le rendement de chaque plan par rapport à la variable fournissant l'information utilisée aux fins d'optimisation et d'équilibrage. Ensuite, l'analyse met principalement l'accent sur l'estimation de  $Y$ .

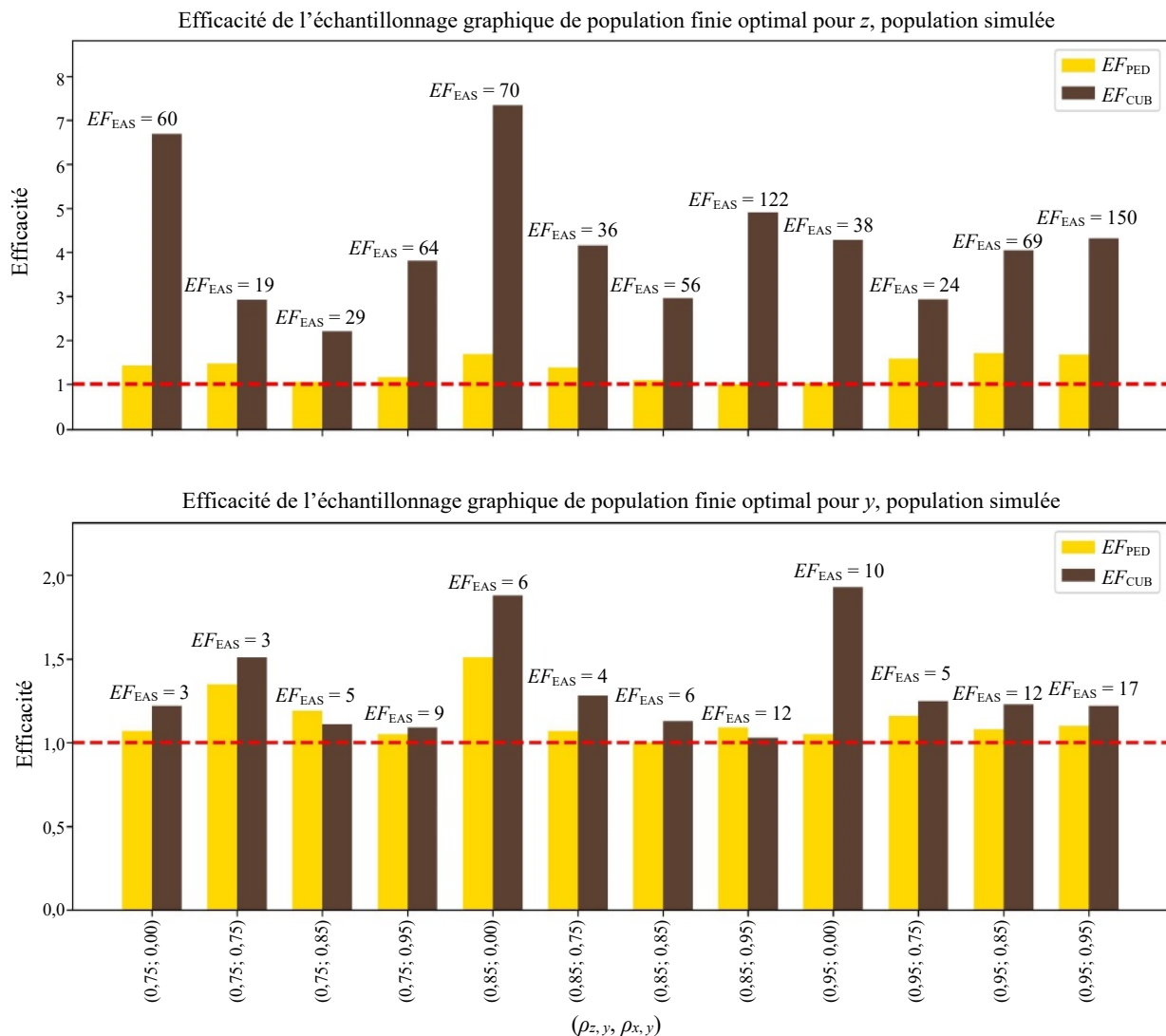
De façon générale, l'EGPF optimal a démontré un bon rendement lorsque  $\rho_{x,y} = 0$ , ce qui correspond au cas d'un échantillonnage avec probabilités égales. Dans ces scénarios, la corrélation entre  $z$  et  $y$  joue un rôle crucial, puisque le gain d'efficacité lors de l'estimation de  $Y$  dépend de la capacité de l'information provenant de  $z$  à appuyer l'estimation de  $Y$ . Lors du passage à un échantillonnage avec probabilités inégales, la situation change : la variation de l'estimateur NHT est alors principalement influencée par des combinaisons linéaires de  $z_k / \pi_k$  pour  $k \in \mathbb{S}$ , ce qui rend la corrélation entre  $z_k / \pi_k$  et  $y_k / \pi_k$  particulièrement importante. Comme cela a été démontré, l'algorithme 5 a su relever efficacement ces défis; l'EGPF optimal a surpassé systématiquement ses concurrents, les méthodes PED et CUB, ou égalé au minimum leur rendement. L'efficacité de l'EGPF optimal par rapport à la méthode EAS est substantielle dans tous les scénarios et, notamment, la méthode PED surpasse la méthode CUB dans presque tous les cas.

Compte tenu de l'ampleur des gains d'efficacité, l'EGPF optimal a affiché un rendement supérieur pour l'estimation du paramètre de la variable auxiliaire utilisée pour l'optimisation,  $z$  (graphique du haut). De plus, en ce qui concerne le nombre de scénarios dans lesquels l'EGPF optimal surpasse strictement ses

concurrents, les méthodes PED et CUB, son rendement était particulièrement remarquable pour estimer le paramètre de la variable principale,  $y$  (graphique du bas).

Dans l'ensemble, ces études par simulation confirment la robustesse et l'efficacité de l'EGPF optimal dans tous les contextes examinés, couvrant diverses combinaisons de  $\rho_{z,y}$  et de  $\rho_{x,y}$ .

**Figure 6.1** Efficacité de l'échantillonnage graphique de population finie optimal par rapport à l'échantillonnage aléatoire simple (EAS), au plan d'échantillonnage déterminantal (PED) et à la méthode du cube (CUB), selon différents scénarios de corrélation dans des populations simulées. Les barres représentent l'efficacité par rapport aux méthodes PED et CUB, tandis que l'efficacité par rapport à l'EAS est indiquée par des valeurs annotées au-dessus des barres pour améliorer la lisibilité, compte tenu des valeurs élevées observées pour l'EAS. La ligne horizontale rouge indique le niveau de référence où l'efficacité est égale à 1. En haut : efficacité pour  $z$ ; en bas : efficacité pour  $y$



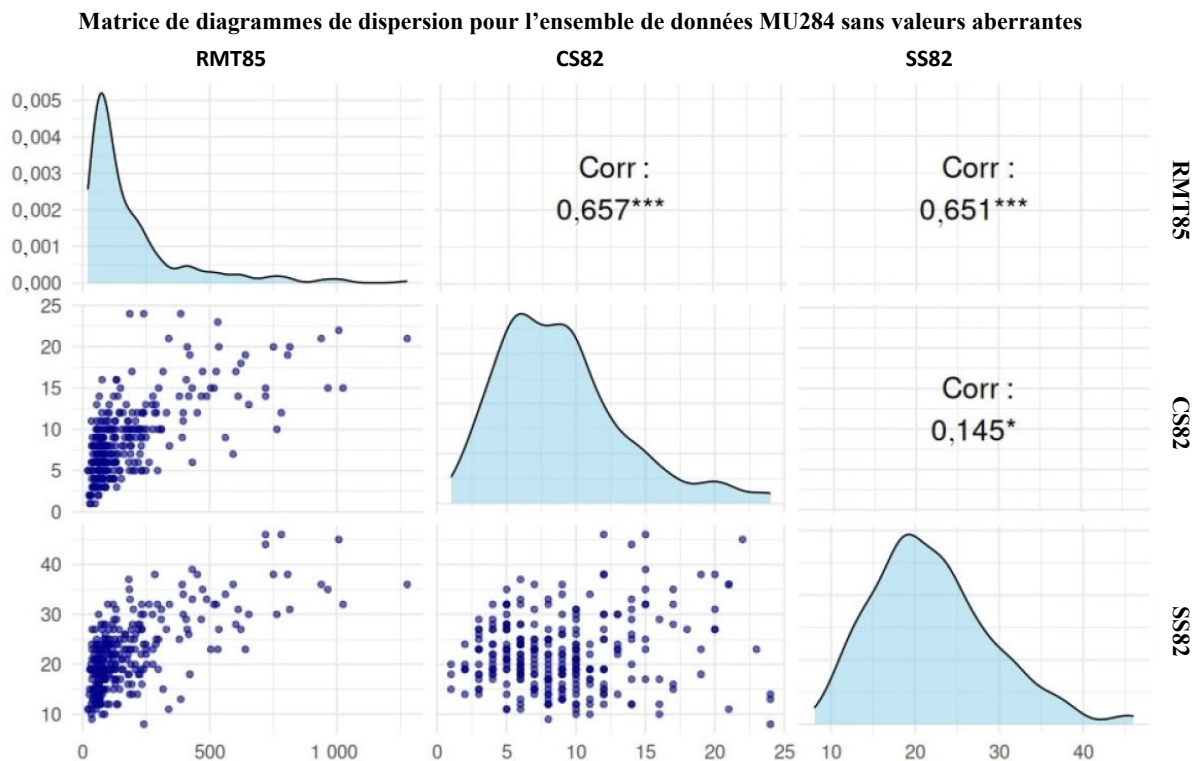
## 6.2 Données de l'ensemble de données MU284

Maintenant que l'algorithme 5 a été évalué dans différents scénarios simulés mettant en jeu des corrélations entre les variables principale et auxiliaires, la présente sous-section vise à examiner son rendement dans une étude de cas réelle reposant sur l'ensemble de données MU284 de Särndal, Swensson et Wretman (1992), qui contient des données sur 284 municipalités suédoises. Pour améliorer la lisibilité des visualisations et pour s'assurer que les résultats puissent être généralisés, trois valeurs aberrantes ont été retirées pour prévenir que les conclusions ne soient influencées par les valeurs extrêmes. L'évaluation porte sur les variables suivantes :

- $y = \text{RMT85}$  : recettes provenant de l'impôt municipal de 1985 (en millions de couronnes);
- $z = \text{CS82}$  : nombre de sièges conservateurs au conseil municipal;
- $x = \text{SS82}$  : nombre de sièges sociaux-démocrates au conseil municipal.

La figure 6.2 présente les distributions et les corrélations des variables sélectionnées. Les variables principale et auxiliaires ont été choisies de manière à ce que l'estimation de  $y$  à partir de  $x$  et de  $z$  soit pertinente compte tenu de la disponibilité temporelle des données. Les variables  $x$  et  $z$  présentaient toutes deux des corrélations modérées avec  $y$  ( $\rho_{z,y} = 0,66$ ;  $\rho_{x,y} = 0,65$ ), ce qui fournissait un contexte réaliste pour évaluer l'efficacité de l'algorithme 5.

**Figure 6.2** Matrice de diagrammes de dispersion et distributions pour les trois variables sélectionnées provenant de l'ensemble de données MU284 après retrait des valeurs aberrantes



**Note :** La figure présente les coefficients de corrélation de Pearson; les niveaux de signification sont indiqués par des symboles : \* pour une valeur de  $p < 0,05$  et \*\*\* pour une valeur de  $p < 0,001$ .

Outre l'examen de l'effet de la corrélation modérée dans un contexte réaliste, le rôle de  $N_{\text{nœud}}$ , un paramètre important de l'algorithme 5, a également été étudié. En général, abstraction faite de la nature aléatoire inhérente au parcours de recherche, des valeurs de  $N_{\text{nœud}}$  plus élevées accroissent l'étendue de la recherche à chaque itération et devraient mener à une plus grande efficacité de l'EGPF optimal. À cette fin, des combinaisons de trois tailles d'échantillon ( $n = 5, 10, 15$ ) et de trois niveaux de  $N_{\text{nœud}}$  (10, 30 et 50) ont été prises en considération. Le nombre d'itérations a été fixé à  $\Lambda = 2\,500$ .

Les résultats sont présentés à la figure 6.3. Comme précédemment, les gains d'efficacité relatifs aux méthodes PED et CUB sont présentés sous forme de diagrammes à barres, tandis que les gains d'efficacité relatifs à l'EAS sont indiqués par des valeurs annotées au-dessus des barres. Les résultats montrent que, comme prévu, une augmentation de  $N_{\text{nœud}}$  améliore généralement l'efficacité de l'EGPF optimal. Pour des valeurs suffisamment élevées de  $N_{\text{nœud}}$ , l'EGPF optimal surpasse les méthodes EAS, PED et CUB dans presque tous les cas, malgré les corrélations modérées entre  $x$  et  $y$ , et entre  $z$  et  $y$ . Dans ce cas-ci, comme pour la population simulée, lorsque l'on considère les cas où l'EGPF optimal surpasse strictement ses concurrents, le scénario plus réaliste d'estimation de  $Y$  présente un rendement plus marqué que celui de l'estimation de  $Z$ , en particulier lorsque  $N_{\text{nœud}}$  a une valeur suffisamment élevée. Toutefois, en ce qui concerne l'ampleur des gains d'efficacité, l'EGPF optimal a affiché un meilleur rendement pour l'estimation de  $Z$ . Cela laisse entendre que l'EGPF optimal est en mesure d'exploiter plus efficacement l'information provenant de  $z$  pendant l'optimisation, ce qui améliore ainsi à la fois l'estimation de  $z$  et de  $y$  plus efficacement que les méthodes concurrentes, PED et CUB.

La taille d'échantillon  $n$  a généralement eu un effet positif sur l'efficacité de l'EGPF optimal lors de l'estimation de  $Z$ , alors que pour la variable principale  $y$ , cet effet n'a pas été observé de manière aussi systématique. Néanmoins, pour presque toutes les tailles d'échantillon, l'EGPF optimal surpasse globalement les autres plans.

Une observation intéressante est que, dans de nombreux cas, lorsque l'algorithme a été arrêté, les gains d'efficacité de l'EGPF optimal continuaient d'augmenter. Cela indique qu'avec une configuration de calcul ayant un meilleur rendement, et un temps de calcul suffisant, il serait possible de déterminer des plans d'EGPF encore plus efficaces.

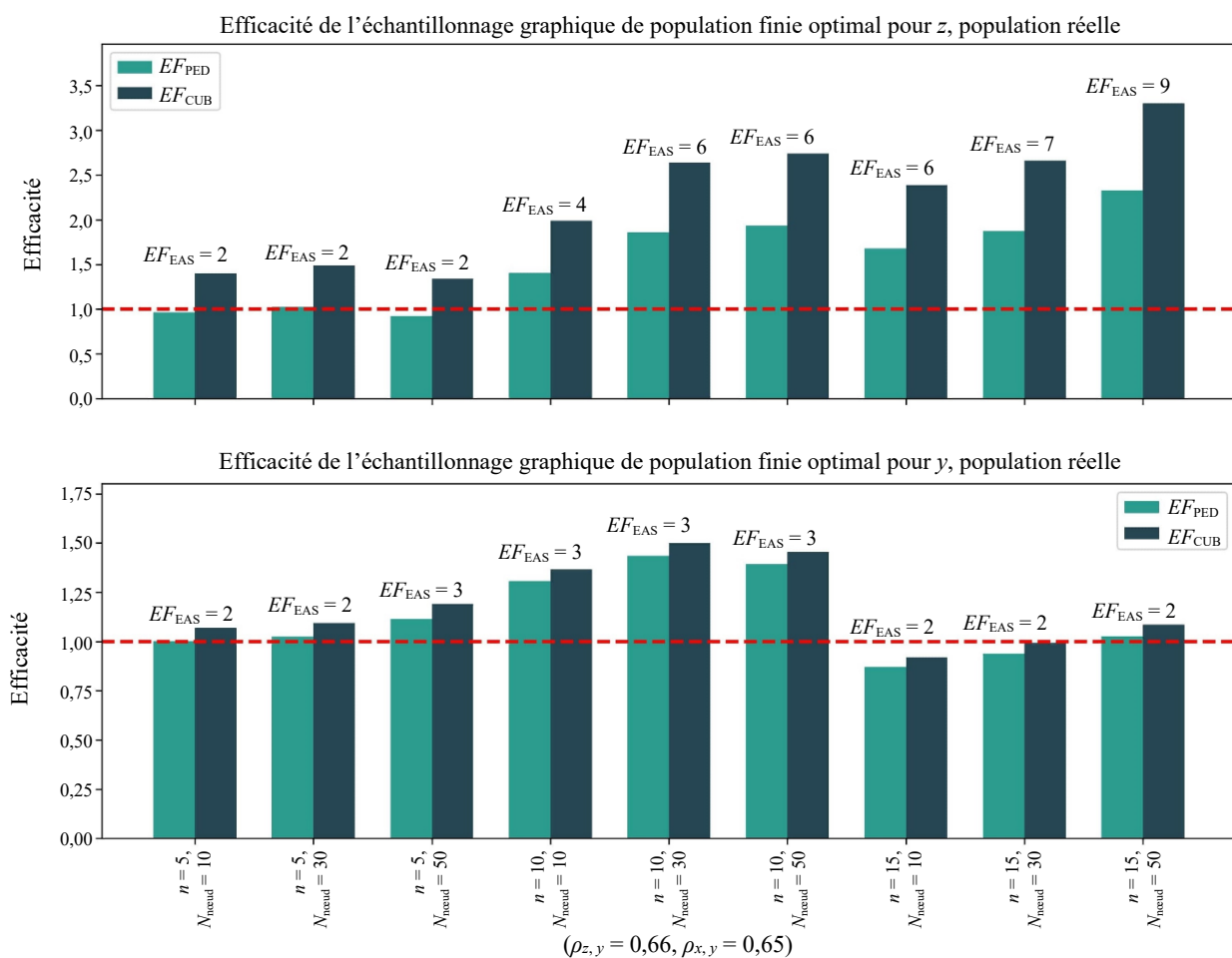
Bien qu'il existe de nombreuses façons d'améliorer davantage l'algorithme 5, l'objectif dans ce cas-ci n'est pas d'en fournir un traitement exhaustif. Voici quelques pistes d'amélioration possibles :

- **Stratégies à démarrages multiples** : lancer la recherche à partir de plusieurs configurations initiales diversifiées afin de réduire la sensibilité aux conditions initiales et d'améliorer l'exploration de l'espace de conception.
- **Élagage adaptatif de l'ensemble ouvert** : élaguer dynamiquement l'ensemble ouvert pour se concentrer sur les plans candidats les plus prometteurs tout en maintenant la diversité.
- **Chocs ou injection aléatoire** : introduire de nouveaux plans initiaux en cours d'optimisation lorsque le processus semble bloqué dans un optimum local, afin de l'aider à s'extraire de régions sous-optimales.

- **Ajustement dynamique de  $N_{\text{nœud}}$**  : faire varier le nombre de plans candidats générés à chaque itération en fonction du comportement de convergence ou de l'état d'avancement de la recherche.
- **Recherche hybride locale-globale** : combiner des étapes basées sur l'algorithme de recherche glouton du meilleur à prime abord avec des méthodes d'exploration stochastique (par exemple échanges ou perturbations aléatoires) afin d'accroître la robustesse de la recherche.

L'étude approfondie de ces stratégies est laissée à des travaux futurs. L'objectif de la présente étude est d'illustrer que les applications de méthodes novatrices d'EGPF, et leur intégration à des algorithmes intelligents, démontrent déjà divers avantages et des gains d'efficacité notables.

**Figure 6.3** Efficacité de l'échantillonnage graphique de population finie optimal par rapport à l'échantillonnage aléatoire simple (EAS), au plan d'échantillonnage déterminantal (PED) et à la méthode du cube (CUB) pour l'ensemble de données MU284, selon différentes valeurs de  $N_{\text{nœud}}$  et tailles d'échantillon. Les barres représentent l'efficacité par rapport aux méthodes PED et CUB, tandis que l'efficacité par rapport à l'EAS est indiquée par des valeurs annotées au-dessus des barres pour améliorer la lisibilité, compte tenu des valeurs élevées observées pour l'EAS. La ligne horizontale rouge indique le niveau de référence où l'efficacité est égale à 1. En haut : efficacité pour  $z$ ; en bas : efficacité pour  $y$



## 7. Résumé et suggestions

L'EGPF présenté dans la présente étude repose sur une approche graphique qui constitue un nouveau pont vers l'échantillonnage de population finie, offrant une nouvelle perspective qui permet de changer la façon classique de raisonner en méthodologie d'échantillonnage. Cette approche novatrice favorise la créativité lors de l'élaboration de nouveaux plans de sondage efficaces en permettant aux chercheurs d'examiner diverses stratégies d'échantillonnage. Par de simples réorganisations de segments, l'EGPF peut générer des plans allant de la construction systématique de référence de Madow (1949), qui entraîne souvent de nombreuses entrées de PIDO nulles, à d'autres solutions caractérisées par des probabilités de plan largement distribuées et une structure de PIDO nettement plus dense. Une telle ampleur d'exploration était auparavant hors de portée des algorithmes mathématiques traditionnels.

Deux composantes notables de ce cadre (l'EGPF chaotique et l'EGPF optimal) illustrent la façon dont l'EGPF peut permettre soit de générer de nouveaux plans, soit de déterminer des configurations optimales en fonction de critères particuliers. De plus, la flexibilité inhérente de l'EGPF se prête naturellement à l'intégration à une grande variété d'algorithmes intelligents. Au-delà de l'algorithme de recherche glouton du meilleur à prime abord utilisé dans le présent article, d'autres stratégies d'optimisation (notamment les algorithmes génétiques, simuler le recuit et l'optimisation par essaim de particules) pourraient être adaptées pour améliorer davantage l'efficacité et la précision.

Malgré ces atouts, l'approche d'EGPF présente un défi notable lorsque la taille de la population ( $N$ ) augmente. En particulier, à mesure que le nombre total d'échantillons possibles ( $T = 2^N$ ) augmente de façon exponentielle, la gestion et la conservation de l'ensemble complet de plans deviennent coûteuses sur le plan des calculs. Néanmoins, compte tenu des progrès récents en calcul à haute performance et du développement continu de technologies d'informatique quantique, ces contraintes deviennent de plus en plus gérables pour les petites et moyennes populations. Pour les populations plus importantes et les scénarios de grandes dimensions, des travaux futurs portant sur des algorithmes intelligents plus sophistiqués ou sur des techniques de représentation compressée pourraient améliorer sensiblement l'évolutivité et l'applicabilité pratique de l'EGPF.

En ce qui concerne l'avenir, une orientation naturelle consiste à formaliser la façon dont d'autres plans standard (échantillonnage aléatoire simple, stratifié, systématique, à plusieurs degrés) s'inscrivent dans cette construction fondée sur des barres, en précisant les contraintes et les règles de mise à jour appropriées. Une deuxième priorité est d'examiner les conjectures formulées précédemment concernant le comportement limitatif des mises à jour chaotiques (avec et sans contrainte de taille fixe), en clarifiant les hypothèses, en établissant des taux de concentration et en évaluant la robustesse aux spécifications de la mise en œuvre.

Grâce à sa flexibilité en matière de réarrangement des segments, le cadre fournit un mécanisme unifié permettant d'examiner et d'optimiser un large éventail de critères de plan, comme la variance ou l'efficacité anticipée, la dispersion spatiale, et la couverture distributionnelle, au sein d'une construction unique. En ce sens, la méthode offre des avenues prometteuses pour la génération de plans d'échantillonnage diversifiés, efficaces et opérationnellement pertinents.

## Remerciements

L'auteur exprime sa sincère reconnaissance à trois de ses étudiants particulièrement talentueux, Mehdi H. Moghadam, Mehdi Mohebbi et Amir Mohammad Hosseini Nasab, pour leurs précieuses contributions et leurs commentaires éclairés, particulièrement à l'étape de la programmation de la recherche. L'auteur est également profondément reconnaissant à Vincent Loonis (Institut national de la statistique et des études économiques de France) pour ses généreux conseils et ses commentaires très utiles, qui ont contribué de manière importante aux études par simulation. Enfin, l'auteur remercie vivement le professeur Yves Tillé pour les discussions enrichissantes auxquelles il a pris part lors de la visite de l'auteur à l'Université de Neuchâtel. Celles-ci l'ont grandement aidé à peaufiner les idées présentées dans l'article.

## Bibliographie

- Brewer, K.R.W. et Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer.
- Deville, J.-C. et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.
- Hansen, M.H. et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- Horvitz, D.G. et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Loonis, V. (2023). [Construire tous les plans de sondage déterminantaux](https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00008-fra.pdf). *Techniques d'enquête*, 49(2), 445-476. Accessible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00008-fra.pdf>.
- 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. *Annals of Mathematical Statistics*, 20, 333-354.
- Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.
- Panahbehagh, B., Mohebbi, M., Hosseini Nasab, A. et Moghadam, M.H. (2025). Graphical-sampling. Accessible à l'adresse : <https://pypi.org/project/graphical-sampling/>.
- Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Thompson, S.K. et Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

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

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations*. Hoboken: Wiley.