Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Commentaires à propos de l’article « Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle »

Jean-Claude Deville : passionné de mathématiques, chercheur de haut vol et visionnaire

par Camelia Goga et Anne Ruiz-Gazen

Date de diffusion : le 3 janvier 2024

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.

Vous pouvez également communiquer avec nous par :

Courriel à 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 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 le ministre responsable de Statistique Canada

© Sa Majesté le Roi du chef du Canada, représenté par le ministre de l'Industrie, 2024

Tous droits réservés. 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.

Techniques d'enquête, décembre 2023
Vol. 49, nº 2, pp. 347-364
Statistique Canada, nº 12-001-X au catalogue

Commentaires à propos de l'article « Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle »

Jean-Claude Deville : passionné de mathématiques, chercheur de haut vol et visionnaire

Camelia Goga et Anne Ruiz-Gazen[1]

Résumé

Jean-Claude Deville compte parmi les plus éminents chercheurs dans la théorie et la pratique des sondages. Ses travaux sur l'échantillonnage équilibré, l'échantillonnage indirect et le calage en particulier sont reconnus au niveau international et largement utilisés en statistique officielle. Il est également pionnier dans le domaine de l'analyse statistique des données fonctionnelles. Le présent article nous donne l'occasion de reconnaître l'immense travail qu'il a accompli, et de lui rendre hommage. Dans la première partie, nous évoquons brièvement la contribution de Jean-Claude à l'analyse statistique en composantes principales fonctionnelles. Nous détaillons également certaines extensions récentes de ses travaux au croisement des domaines de l'analyse statistique des données fonctionnelles et de la théorie des sondages. Dans la seconde partie, nous présentons une extension de son travail dans le domaine de l'échantillonnage indirect. Ces résultats de recherche sont motivés par des applications concrètes et illustrent l'influence de Jean-Claude sur notre travail de chercheuses.

Mots-clés : Estimation de la consommation d'électricité; analyse statistique des données fonctionnelles; méthode généralisée du partage des poids; échantillonnage indirect; estimation du trafic postal.

1. Introduction

Au cours de sa longue carrière d'« administrateur » à l'Insee (Institut national de la statistique et des études économiques), l'organisme national de statistique en France, Jean-Claude Deville a réalisé de nombreuses études statistiques et relevé de nombreux défis en théorie des sondages. L'une de ses plus grandes préoccupations était de résoudre les problèmes des praticiens, comme il l'avait lui-même affirmé dans l'introduction de Deville (1974) : « L'article qui suit est né d'un problème tout à fait concret : l'étude de la constitution progressive des familles, du calendrier des naissances en fonction de la durée de mariage. » Il n'hésitait pas, par exemple, à aller sur le terrain et à participer au tirage d'échantillons. Un jour, à 5 h du matin, il s'est présenté dans un centre de tri pour assister au tirage de lettres effectué par le service postal français. À cette occasion, il a dit : « Nous pouvons sentir vivre la statistique! » La source de cette citation est une communication personnelle d'Alain Dessertaine, ingénieur de recherche à La Poste (le service postal français) qui a collaboré avec Jean-Claude Deville sur des enquêtes menées par son organisme (voir aussi la section 3 du présent article). Le goût de Jean-Claude Deville pour les applications, combiné à son extraordinaire culture et à sa passion pour les mathématiques, ont donné naissance à de nouvelles théories

1. Camelia Goga, LMB, Université de Franche-Comté. Courriel : camelia.goga@univ-fcomte.fr; Anne Ruiz-Gazen, EET, Université Toulouse Capitole. Courriel : anne.ruiz-gazen@tse-fr.eu.

et méthodes de statistique, en particulier dans le domaine des sondages. Ces méthodes ont notamment apporté des solutions concrètes aux problèmes initiaux et même au-délà.

Notre collaboration avec Jean-Claude a été de longue durée (depuis le début des années 2000). Nous avons assisté à un grand nombre de ses fascinants exposés et de ses cours, nous avons discuté longuement et passionnément de mathématiques avec lui (parfois sur des bouts de papier) et nous avons publié plusieurs articles avec lui. Il a grandement influé sur nos méthodes de travail en tant que chercheuses en nous incitant à partir de problèmes appliqués et à concevoir des résultats théoriques généraux, alors même que nous ne travaillions pas dans le domaine de la statistique officielle. Il était des plus inspirants et nous a transmis sa passion de la théorie des sondages, et plus généralement, de la statistique et des mathématiques, mais aussi son goût pour les applications. Dans ce qui suit, nous allons illustrer la façon dont Jean-Claude a exercé une influence sur notre travail et, ce faisant, sur le travail de nos doctorants. À cette fin, nous examinerons deux applications concrètes que nous avons rencontrées dans notre carrière, à savoir l'estimation des courbes de consommation d'électricité et du trafic postal. Ces applications nous ont permis de faire quelques (petites) avancées dans la théorie des sondages qui sont des prolongements du travail de Jean-Claude dans des domaines où il a été un précurseur : l'analyse statistique des données fonctionnelles et l'échantillonnage indirect.

La section 2 de notre discussion est consacrée au travail de Jean-Claude sur les données indépendantes de type fonctionnel, développé durant les années 1970 avant qu'il ne s'intéresse à la théorie des sondages. On donne également dans cette section une extension de l'analyse statistique des données fonctionnelles au cas des données dépendantes issues des enquêtes par sondage dans une population de courbes, telle que la population des courbes de consommations électriques dans le but d'estimer la courbe de consommation électrique totale. Dans la section 3, nous décrivons un problème d'échantillonnage indirect rencontré en pratique à La Poste. Nous nous concentrons sur un type particulier de structure de liens entre la population échantillonnée et la population cible, appelé « tous pour un », et nous montrons comment il est possible d'étendre les résultats de Deville et Lavallée (2006). Nous considérons également un échantillonnage indirect double particulier où le nombre de liens à observer peut être considérablement réduit. À la section 4, enfin, nous concluons avec quelques réflexions personnelles sur notre collaboration avec Jean-Claude.

2. Deville et l'analyse statistique des données fonctionnelles : un visionnaire des mégadonnées

2.1 Données fonctionnelles indépendantes

L'article pionnier de 1974, « *Méthodes statistiques et numériques de l'analyse harmonique* », publié malheureusement uniquement en français, développe un nouveau type de statistique, la statistique des données fonctionnelles, largement popularisée plus tard par l'ouvrage de Ramsay et Silverman (2005). L'article de 1974 de J.C. Deville porte sur une étude réalisée en 1962 sur 240 000 femmes âgées de moins de 70 ans dans le but principal de reconstituer individuellement le calendrier de la formation de chaque famille. La date de naissance de chaque femme ainsi que les dates de naissance de ses enfants ont été renseignées, tout comme le niveau de scolarité, la profession et la position géographique de la famille cette

année-là, et d'autres renseignements sur le mari, etc. Dans Deville (1977), article qui détaille l'application de la méthode proposée dans Deville (1974) aux données de la constitution des familles, Deville affirme que « devant des données d'une telle richesse le statisticien éprouve une certaine perplexité ». Deville était à l'aube de l'analyse des mégadonnées et l'utilisation de méthodes classiques comme l'analyse en composantes principales (ACP) menait à des résultats impossibles à interpréter. Des relevés statistiques basés sur une division plus fine du temps, en trimestres plutôt qu'en années, ont fait penser à Deville que « le caractère de paramètre continu du temps n'était pas sans influence sur le choix d'une méthode d'analyse » (Deville, 1974, chapitre 1). Il a considéré alors que chaque individu k était caractérisé par une courbe temporelle, fonction $Y_k(t)$ pour t variant dans un intervalle fermé $[0, \mathcal{T}]$ « au lieu d'être caractérisé par un vecteur de dimension finie » (Deville, 1977) et il a étendu l'analyse en composantes principales à ce nouveau cadre fonctionnel. Dans l'étude des calendriers des formations des familles, 0 représente l'année de mariage et $\mathcal{T} = 20,$ les années écoulées depuis l'année du mariage. Deville considère cette période postérieure à la date de mariage comme le laps de temps maximal pour avoir des enfants, puisque très peu d'enfants naissent après 20 ans de mariage. La valeur $Y_k(t)$ représente le nombre d'enfants à l'année t (depuis l'année du mariage). L'étude a porté sur seulement 100 000 familles considérées comme *complètes*, c'est-à-dire non dissoutes après 20 ans de mariage et pour lesquelles l'épouse était âgée de moins de 45 ans au moment du mariage.

Les observations $Y_k = (Y_k(t))_{t\in[0,\mathcal{T}]}$ sont des courbes ou des fonctions de t et le traitement de tels objets exige des outils de la théorie des processus aléatoires et de l'analyse fonctionnelle. Des objectifs nouveaux et différents apparaissent dans ce nouveau cadre fonctionnel et, dans son papier novateur, Deville a posé les bases d'un nouveau type de statistique appelée, plusieurs années après, *analyse statistique des données fonctionnelles* (Ramsay et Silverman, 2005). Dans cet espace de dimension infinie, une première étape consiste à mieux représenter les données dans un espace de petite dimension qui permet de décrire les données fonctionnelles et de faciliter leur interprétation. Pour ce faire, Deville (1974) a supposé que les Y_k sont des fonctions aléatoires indépendantes appartenant à l'espace de Hilbert $L^2[0, \mathcal{T}],$ l'espace des fonctions de carré intégrable définies dans l'intervalle fermé $[0, \mathcal{T}],$ muni du produit scalaire $< f, g > = \int_0^{\mathcal{T}} f(t)\, g(t)\, dt$ et de la norme induite

$$\|f\| = \left(\int_0^{\mathcal{T}} f^2(t)\, d(t)\right)^{1/2}$$

pour $f,\ g \in L^2[0, \mathcal{T}].$ Il a démontré que Y_k pouvait être représenté de la manière suivante :

$$Y_k(t) = \mu(t) + \sum_{j\geq 1} < Y_k - \mu, v_j > v_j(t), \quad t \in [0, \mathcal{T}],$$

où $\mu(t)$ est la moyenne de $Y_k(t)$ (c'est-à-dire le nombre moyen d'enfants des familles à l'étude à l'instant t), où $v_j(\cdot),\ j \geq 1$ sont des fonctions orthonormées de $L^2[0, \mathcal{T}]$ qui dépendent de $t,$ mais non de l'individu $k,$ et qui sont appelées *les harmoniques du processus* dans Deville (1974) par analogie avec l'analyse harmonique des signaux périodiques/stationnaires dans des bases de Fourier. Les composantes $< Y_k - \mu, v_j >,\ j \geq 1$ dépendent de l'individu $k,$ mais sont indépendantes du temps, elles sont non-corrélées entre elles et de variance égale à $\lambda_j,$ avec $\lambda_1 \geq \lambda_2 \geq \ldots.$ Les fonctions Y_k peuvent être approximées dans un espace de dimension plus petite q comme suit :

$$Y_k(t) \;=\; \mu(t) + \sum_{j=1}^{q} < Y_k - \mu, v_j > v_j(t) + R_{q,k}(t), \quad t \in [0, \mathcal{T}], \tag{2.1}$$

où le reste $R_{q,k}(t)$ est le plus petit possible selon un critère de variance intégrée. La relation (2.1) nous permet de représenter le mieux possible dans un espace de petite dimension q les variations des courbes Y_k par rapport à la courbe moyenne μ. L'expression (2.1) est l'extension fonctionnelle du résultat classique de la décomposition sur les q premières composantes principales; Deville (1974, chapitre 4) a considéré cette décomposition comme une *analyse en composantes principales fonctionnelle* (ACPF), terminologie utilisée plus tard par un grand nombre d'auteurs (Ramsay et Silverman, 2005). Les fonctions $v_j(\cdot)$ et les quantités $\lambda_j, j \geq 1$ sont étroitement liées à l'opérateur de covariance $\mathbf{\Gamma}$ de Y_k, qui est l'équivalent fonctionnel de la matrice des covariances. L'opérateur de covariance $\mathbf{\Gamma}$ est défini sur l'espace $L^2[0, \mathcal{T}]$ dans $L^2[0, \mathcal{T}]$ par :

$$\mathbf{\Gamma} a(r) \;=\; \int_0^{\mathcal{T}} \gamma(r,t)\, a(t)\, dt, \tag{2.2}$$

pour toute fonction $a \in L^2[0, \mathcal{T}]$ et $\gamma(r,t) = \mathrm{cov}(Y_k(r), Y_k(t))$, la covariance entre $Y_k(r)$ et $Y_k(t)$ pour $r, t \in [0, \mathcal{T}]$. Les quantités λ_j et $v_j(\cdot)$ sont respectivement les valeurs propres et les fonctions propres de $\mathbf{\Gamma}$:

$$\mathbf{\Gamma} v_j(t) = \lambda_j v_j(t), \tag{2.3}$$

avec $\lambda_j \geq 0$, $j \geq 1$, classés par ordre décroissant $\lambda_1 \geq \lambda_2 \geq \ldots$. Dans un espace de dimension p, $\mathbf{\Gamma}$ est la matrice de covariance classique de taille $p \times p$; λ_j et v_j sont les valeurs propres et les vecteurs propres habituels. Le lecteur familier avec l'analyse en composantes principales multivariée (Jolliffe, 2002), reconnaîtra dans $< Y_k - \mu, v_j >$ les scores de la j^{e} composante principale de variance égale à λ_j. Il interprétera les valeurs propres comme la partie de la variance totale qui est expliquée par la j^{e} composante principale. La même interprétation de λ_j est valable dans ce nouveau cadre fonctionnel (Deville, 1974). L'espace de dimension q engendré par les vecteurs propres $v_j, j = 1, \ldots, q$ représente les principaux modes de variation autour de la courbe moyenne μ des données tout au long du temps t. On trouve dans Deville (1977) l'application de cette nouvelle approche statistique aux données des trajectoires familiales ainsi qu'une analyse approfondie de différentes typologies de familles mises en évidence par l'analyse en composantes fonctionnelles.

2.2 Données fonctionnelles dépendantes issues des enquêtes par sondage

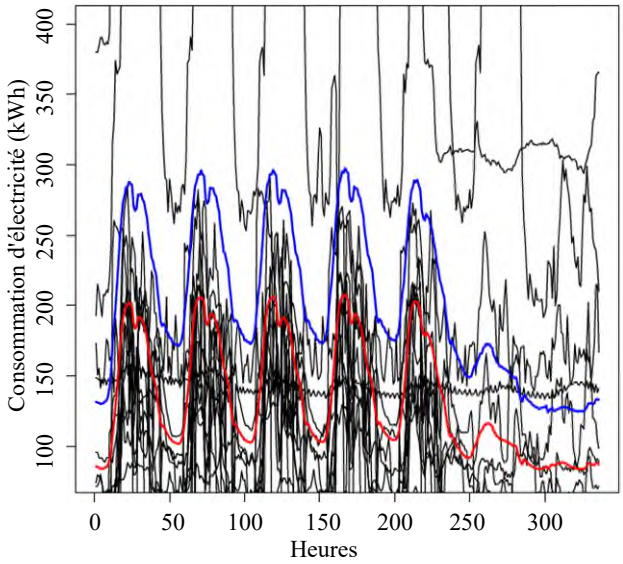
La méthode conçue dans Deville (1974) est applicable à divers contextes et Deville en a donné plusieurs exemples au chapitre 8 du même article. Il a eu des idées innovantes comme l'attestent les études faisant appel à ces outils fonctionnels qui ont fait leur apparition dans divers domaines comme la chimiométrie (Hastie et Mallows, 1993), l'économie (Kneip et Utikal, 2001), la climatologie (Besse, Cardot et Stephenson, 2000) et la biologie (Chiou, Müller et Wang, 2003), pour n'en citer que quelques articles parmi les nombreuses références dans ce domaine.

Une nouvelle ère statistique est apparue avec les progrès des modes automatisés de collecte de données et l'augmentation des capacités de stockage. Grâce aux capteurs intelligents, il est désormais très courant

de disposer des données très volumineuses. Les courbes de charge d'électricité constituent un tel exemple. Plus précisément, Électricité de France (EDF) a installé plus de 30 millions de compteurs intelligents dans presque tous les ménages et les entreprises. Ces compteurs sont capables de mesurer et de transmettre la consommation individuelle d'électricité à des pas de temps très fins (toutes les 30 minutes, voire à chaque minute ou seconde). Comme le pas de discrétisation est très fin, les unités statistiques peuvent être considérées comme des fonctions du temps. Considérons, par exemple, une population test de $N = 18\,902$ entreprises françaises dont la consommation d'électricité a été mesurée toutes les demi-heures sur une période de deux semaines. Nous présentons dans la figure 2.1 un échantillon de 20 courbes de charge électrique extraites de cette population test ainsi que les profils d'électricité moyen et médian calculés sur la population entière. Dans cet exemple, $Y_k(t)$ est la consommation d'électricité de chaque k à l'instant t.

Néanmoins, en raison des contraintes techniques et budgétaires dues à une bande passante restreinte ou aux coûts de stockage de vastes bases de données, l'analyse de l'ensemble des données produites peut être très difficile ou très coûteuse. Chiky (2009) a montré que, si nous nous intéressons uniquement à des indicateurs simples tels que les trajectoires totale ou moyenne, l'utilisation des techniques d'échantillonnage, même très simples comme l'échantillonnage aléatoire simple sans remise, constituent une attrayante alternative aux techniques de compression de signal, puisqu'elles permettent d'obtenir des estimations précises à un prix raisonnable. En se basant sur ces résultats, EDF a envisagé d'utiliser des stratégies d'échantillonnage efficaces pour estimer la consommation totale d'électricité, ce qui a fait naître un certain nombre de travaux de recherche combinant l'analyse statistique de données fonctionnelles et la théorie de l'échantillonnage. Une collaboration de recherche a été mise en place entre ERDF (Alain Dessertaine) et l'Université de Bourgogne (Hervé Cardot et Camelia Goga) et plusieurs thèses de doctorat ont été consacrées à des sujets liés à ces thématiques (Étienne Josserand, Pauline Lardin-Puech, Anne De Moliner).

Figure 2.1 Échantillon de 20 courbes de consommation d'électricité mesurée toutes les demi-heures sur une période d'une semaine.



La courbe de consommation moyenne dans la population figure en bleu gras et la courbe de consommation médiane, en rouge.

Le premier travail combinant les données fonctionnelles et les techniques d'échantillonnage a été proposé dans Cardot et coll. (2010), qui ont étudié l'analyse en composantes principales fonctionnelles (ACPF) pour des données issues des enquêtes par sondage. Plusieurs articles ont traité ensuite l'estimation de la courbe totale ou moyenne avec des plans de sondage (Cardot et Josserand, 2011; Cardot, Degras et Josserand, 2013). Une problématique importante avec ce nouveau type de données est comment construire des bandes de confiance asymptotiques de taux de couverture contrôlés (Cardot et Josserand, 2011). Cardot, Dessertaine, Goga, Josserand et Lardin (2013) ont comparé, du point de vue de la précision des estimateurs de la courbe moyenne d'électricité, plusieurs approches qui considèrent l'information auxiliaire. Ils ont aussi comparé les largeurs de bandes de confiance. La conclusion de leur étude empirique est que, en intégrant de l'information auxiliaire lors de l'étape d'échantillonnage ou de l'étape de l'estimation, on améliore grandement l'efficacité des estimateurs. En particulier, la largeur des bandes peut être nettement réduite. La justification théorique de ces résultats a été établie par Cardot, Goga et Lardin (2013) et par Cardot, Goga et Lardin (2014).

Nous présentons brièvement ci-après l'analyse en composantes principales fonctionnelles pour des données issues des enquêtes par sondage. Ce travail étend les résultats de Deville (1974) à des données fonctionnelles non indépendantes ainsi que l'approche par fonction d'influence de Deville (1999).

2.2.1 Analyse en composantes principales fonctionnelles pour des données issues d'enquête par sondage

Considérons une population finie U de taille N, sur laquelle nous observons, pour chaque unité k, une fonction déterministe de temps $Y_k = (Y_k(t))_{t\in[0,\mathcal{T}]}$. La courbe totale est définie par $t_Y = \sum_{k\in U} Y_k$ et la trajectoire moyenne par $\mu_N = t_Y / N$. La valeur de t_Y ou μ_N au point de mesure $t \in [0,\mathcal{T}]$ s'obtient directement comme $t_Y(t) = \sum_{k\in U} Y_k(t)$ et $\mu_N(t) = \sum_{k\in U} Y_k(t)/N$ respectivement. La figure 2.1 présente la courbe de consommation moyenne d'électricité (en bleu) calculée sur la population d'intérêt.

Pour effectuer l'analyse en composantes principales fonctionnelles (ACPF), nous devons d'abord estimer la fonction de covariance au niveau de la population. Pour r et t dans $[0,\mathcal{T}]$, la fonction de covariance $\gamma(r,t)$ entre $(Y_k(r))_{k\in U}$ et $(Y_k(t))_{k\in U}$ est donnée par :

$$\gamma(r,t) = \frac{1}{N}\sum_{k\in U}(Y_k(r)-\mu_N(r))\,(Y_k(t)-\mu_N(t)), \quad (r,t)\in[0,\mathcal{T}]\times[0,\mathcal{T}],$$

et l'opérateur de covariance $\mathbf{\Gamma}$ associé est donné en (2.2). Cet opérateur de covariance a les expressions équivalentes suivantes :

$$\mathbf{\Gamma} = \frac{1}{N}\sum_{k\in U}(Y_k-\mu_N)\otimes(Y_k-\mu_N) = \frac{1}{N}\sum_{k\in U} Y_k\otimes Y_k - \mu_N\otimes\mu_N,$$

où le produit tensoriel de deux éléments a et b de $L^2[0,\mathcal{T}]$ est défini par :

$$a\otimes b(y) = <a, y> b, \text{ pour tout } y\in L^2[0,\mathcal{T}].$$

Enfin, les éléments propres de $\mathbf{\Gamma}$ sont donnés par :

$$\mathbf{\Gamma} v_j(t) = \lambda_j v_j(t), \quad t \in [0, \mathcal{T}], \quad j = 1, \ldots, N. \tag{2.4}$$

2.2.2 Estimateurs de l'analyse en composantes principales fonctionnelles fondés sur le plan

La trajectoire moyenne μ_N ou l'opérateur de variance $\mathbf{\Gamma}$, ainsi que les éléments propres sont inconnus parce que nous n'avons pas accès à toutes les unités k de la population U. Ces quantités sont estimées à l'aide d'un échantillon s tiré dans U selon un plan de sondage $p(\cdot)$, avec des probabilités d'inclusion du premier ordre $\pi_k, k \in U$. Skinner, Holmes et Smith (1986) ont étudié un certain nombre de propriétés de l'analyse en composantes principales (ACP) multivariée dans le cadre des données d'enquête et Deville (1999) a déterminé la variance asymptotique des éléments propres d'une ACP en adoptant l'approche par la fonction d'influence. Tous ces paramètres d'intérêt sont des fonctions non linéaires du total t_Y. Les valeurs propres et les fonctions propres de $\mathbf{\Gamma}$ sont définies par l'équation implicite (2.3) et sont également des fonctions non linéaires. L'approche proposée par Deville (1999) consiste à écrire le paramètre d'intérêt comme une fonctionnelle T de la mesure discrète M définie sur l'espace $L^2[0, \mathcal{T}]$:

$$M = \sum_{k \in U} \delta_{Y_k},$$

où δ_{Y_k} est la mesure de Dirac à Y_k avec $k \in U$. Comme décrit dans Cardot, Chaouch, Goga et Labruère, (2008, 2010), μ_N et $\mathbf{\Gamma}$ peuvent être écrites comme fonctionnelles de M :

$$\mu_N = \frac{\int Y dM}{\int dM}, \tag{2.5}$$

$$\mathbf{\Gamma} = \frac{\int (Y - \mu_N) \otimes (Y - \mu_N)\, dM}{\int dM}. \tag{2.6}$$

Les éléments propres de $\mathbf{\Gamma}$, définis par la fonction implicite (2.4), sont aussi des fonctionnelles de M. La mesure M peut être estimée par :

$$\hat{M} = \sum_{k \in s} \frac{1}{\pi_k} \delta_{Y_k}$$

et les estimateurs de μ_N et $\mathbf{\Gamma}$ s'obtiennent en remplaçant M par $\hat{M}$ dans (2.5) et (2.6), ce qui donne :

$$\hat{\mu}_{\text{Haj}} = \frac{\hat{t}_Y}{\hat{N}}, \tag{2.7}$$

où $\hat{N} = \sum_{k \in s} 1/\pi_k$ est l'estimateur de Horvitz-Thompson de N. L'opérateur de variance $\mathbf{\Gamma}$ est estimé par

$$\hat{\mathbf{\Gamma}} = \frac{1}{\hat{N}} \sum_{k \in s} \frac{(Y_k - \hat{\mu}_{\text{Haj}}) \otimes (Y_k - \hat{\mu}_{\text{Haj}})}{\pi_k} = \frac{1}{\hat{N}} \sum_{k \in s} \frac{Y_k \otimes Y_k}{\pi_k} - \hat{\mu}_{\text{Haj}} \otimes \hat{\mu}_{\text{Haj}}. \tag{2.8}$$

Les estimateurs $\hat{\lambda}_j$, de λ_j, et $\hat{v}_j$, de v_j, sont les éléments propres de $\hat{\mathbf{\Gamma}}$, soit

$$\hat{\mathbf{\Gamma}}\hat{v}_j(t) = \hat{\lambda}_j\hat{v}_j(t), \quad t \in [0, \mathcal{T}], \quad j = 1, \ldots, N. \tag{2.9}$$

2.2.3 Approximation asymptotique et estimateurs de la variance

Tous les paramètres d'intérêt de la population peuvent s'écrire comme $T(M)$, où T est une fonctionnelle de degré 0, soit $T(M/N) = T(M)$. Sous des hypothèses faibles sur le plan de sondage et sur les fonctions Y_k, Cardot et coll. (2010) démontrent que les estimateurs $T(\hat{M})$, basés sur le plan de sondage obtenus par les équations (2.7) à (2.9) sont asymptotiquement sans biais pour $T(M)$. Pour déduire la variance asymptotique de $T(\hat{M})$, Cardot et coll. (2010) donnent un développement de von Mises de $T(\hat{M}/N) = T(\hat{M})$ autour de M/N, comme suit :

$$\begin{aligned} T(\hat{M}) - T(M) &= \int IT\left(\frac{M}{N}, Y_k\right) d\left(\frac{\hat{M}}{N} - \frac{M}{N}\right) + R_T \\ &= \sum_{k \in U} IT(M, Y_k)\left(\frac{I_k}{\pi_k} - 1\right) + R_T, \end{aligned}$$

où $I_k = 1$ si l'unité k est sélectionnée dans un échantillon et zéro sinon; $IT(M/N, Y_k) = N \cdot IT(M, Y_k)$, où $IT(M, Y_k)$ est la fonction d'influence de T au point Y_k comme définie dans Deville (1999) et appelée la *variable linéarisée* de T, et R_T est le terme du reste. La fonction d'influence de μ_N est $I\mu_N(M, Y_k) = (Y_k - \mu_N)/N$ (Deville, 1999). En utilisant des arguments de la théorie de la perturbation, Cardot et coll. (2010) montrent que $I\mathbf{\Gamma}(M, Y_k) = \left((Y - \mu_N) \otimes (Y - \mu_N) - \mathbf{\Gamma}\right)/N$ et que

$$\begin{aligned} I\lambda_j(M, Y_k) &= \frac{1}{N}(< Y_k - \mu_N, v_j >^2 - \lambda_j) \\ Iv_j(M, Y_k) &= \frac{1}{N}\sum_{\ell \neq j} \frac{< Y_k - \mu_N, v_j >< Y_k - \mu_N, v_\ell >}{\lambda_j - \lambda_\ell} v_\ell, \end{aligned}$$

en supposant que $\lambda_j \neq \lambda_\ell$ pour tout $j \neq \ell$. Les mêmes expressions ont été obtenues pour $I\lambda_j$ et Iv_j par Deville (1999) pour le cas non fonctionnel. Cardot et coll. (2010) montrent que $R_T = o_p(n^{-1/2})$ pour T donné dans les équations (2.7) à (2.9), et déterminent la variance asymptotique de $T(\hat{M})$. Cardot et coll. (2010) donnent aussi des estimateurs de variance qui sont convergents sous des hypothèses supplémentaires sur les probabilités d'inclusion d'ordre supérieur.

2.2.4 Pourquoi utiliser l'analyse en composantes principales fonctionnelles ?

Les scores de l'analyse en composantes principales, $< Y_k - \mu_N, v_j >$ pour $j = 1, \ldots, q$, indiquent l'écart de la courbe Y_k par rapport à sa fonction moyenne μ_N et permettent d'obtenir une bonne reconstitution des fonctions $Y_k, k \in U$ dans un espace de dimension faible q. Lorsque des variables auxiliaires sont disponibles, Cardot et coll. (2010) proposent de s'en servir pour améliorer l'estimation de t_Y. L'analyse en

composantes principales fonctionnelles (ACPF) peut aussi être utilisée lorsque l'information auxiliaire est de très grande taille; par exemple, la consommation d'électricité relevée toutes les 30 minutes pendant une certaine période peut servir comme information auxiliaire pour construire des estimateurs de t_Y assistés par un modèle ou par calage (Cardot, Goga et Shehzad, 2017). De telles variables auxiliaires sont fonctionnelles, et l'ACPF nous permet de déterminer de nouvelles variables non-corrélées qui peuvent être utilisées ensuite comme des variables de calage (Deville et Särndal, 1992) pour améliorer l'estimation de t_Y, comme suggéré dans Cardot et coll. (2017).

3. Revue et prolongement des idées de Deville sur l'échantillonnage indirect

La question traitée dans cette section nous a été présentée par Alain Dessertaine et Pauline Puech, de La Poste. Ils étaient préoccupés par la perte de précision des estimateurs nationaux du trafic postal après le changement du plan de sondage. Pour mieux comprendre leur problème, nous avons travaillé sur la théorie avec une doctorante, Estelle Medous, et avons fait des progrès intéressants sur le plan méthodologique. Comme il est expliqué à la section 4 dans Ardilly, Haziza, Lavallée et Tillé (2023), Jean-Claude Deville a joué un rôle clé dans l'identification et la dérivation des principes généraux de l'échantillonnage indirect ainsi que des propriétés de ce plan. Les principaux résultats figurent dans Deville et Lavallée (2006) et dans Lavallée (2007). Dans ce qui suit, nous présentons certains prolongements des travaux de recherche de Jean-Claude Deville sur l'échantillonnage indirect tels que détaillés dans Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine et Puech (2023) et Medous (2023).

3.1 Le problème d'échantillonnage de La Poste

La mesure du trafic postal mensuel en France est importante en particulier pour le suivi comptable à La Poste. Toutefois, seule une partie du trafic postal passe par le traitement automatisé. Le trafic postal mensuel doit donc être estimé à l'aide d'une enquête probabiliste. En 1994, cette enquête a été baptisée SYCI (ou Système de collecte de l'information). Il s'agissait de tirer des échantillons des populations de tournées ou itinéraires de postiers en appliquant un plan de sondage stratifié à deux degrés et équilibré, et en utilisant des estimateurs calés. Jean-Claude Deville, alors « chef des méthodes statistiques » à l'Insee, a participé à la validation du cadre méthodologique de cette enquête et a étudié en particulier la précision des estimateurs.

Depuis 2008, l'organisation des tournées des postiers a changé et n'est plus stable dans le temps. Il est devenu impossible de tirer directement un échantillon de la population cible et le plan d'échantillonnage a évolué vers un plan d'échantillonnage indirect appelé SYCI2, pour lequel la population cible est toujours la population de tournées des postiers, mais la population échantillonnée ou frame est celle des adresses postales. Comme il est détaillé dans section 4 de Ardilly et coll. (2023) sur l'échantillonnage indirect, dès que les poids associés aux liens entre les deux populations sont standardisés (leur somme est égale à un), les estimateurs construits selon la méthode généralisée de partage des poids (MGPP) sont sans biais pour le

paramètre d'intérêt. Mais il n'est pas facile de comparer leur précision à celle des estimateurs de l'échantillonnage direct. À La Poste, l'échantillonnage indirect de SYCI2, contrairement à la méthode antérieure d'échantillonnage direct de SYCI, a causé une perte de précision des estimateurs et les écarts-types estimés ont augmenté d'un facteur de 2 à 3. L'étude, résumée ci-dessous, visait à comprendre en profondeur cette perte de précision.

La population frame est formée des adresses postales, et la population cible des tournées des postiers. La structure des liens entre ces deux populations est d'un type particulier car chaque adresse appartient à une seule et unique tournée de postier. De tels liens sont qualifiés de « tous pour un » dans Deville et Lavallée (2006). Nous emploierons l'expression « tous pour un » à partir de maintenant. Comme précisé ci-dessous, pour ce type de structure de liens, il est possible d'aller au-delà de Deville et Lavallée (2006) dans l'étude théorique des estimateurs MGPP.

3.2 Échantillonnage indirect « tous pour un »

Considérons comme paramètres de l'étude les totaux calculés sur la population cible U_T. Comme Deville et Lavallée (2006), nous supposons qu'une base de sondage existe pour une population dite frame U_F, liée à U_T de sorte que toute unité de U_T soit en liaison avec au moins une unité de U_F. De plus, nous considérons la situation de liens « tous pour un » où chaque unité de U_F est liée à une seule et unique unité de U_T. Soit y la variable d'intérêt mesurée sur U_T et soit y_k sa valeur pour la k^e unité dans U_T. Nous désirons estimer $t_y = \sum_{k\in U_T} y_k$, le total de y sur U_T. Nous tirons un échantillon s_F dans U_F selon un plan de sondage $p_F(\cdot)$. Nous pouvons associer à s_F le vecteur $(I_1,\ldots,I_{N_F})'$, où I_i est l'indicateur d'appartenance à un échantillon de l'individu i de U_F, défini par $I_i = 1$ si i est sélectionné et $I_i = 0$ autrement. Nous désignons par $\pi_i = \Pr(i \in s_F)$ la probabilité d'inclusion du premier ordre de l'unité i et par $\pi_{ii'} = \Pr(i, i' \in s_F)$ la probabilité d'inclusion du second ordre des unités i et i'. Nous supposons que toutes les unités i ont une probabilité strictement positive d'inclusion $\pi_i > 0$ et nous désignons par $d_i = 1/\pi_i$ leurs poids d'échantillonnage. L'échantillon s_F de U_F conduit à un échantillon s_T de U_T, qui contient les unités de U_T en liaison avec au moins une unité de s_F.

Considérons, pour tout lien entre les individus $i \in U_F$ et $k \in U_T$, un poids non négatif θ_{ik} tel que θ_{ik} soit positif quand $i \in U_F$ et $k \in U_T$ sont en liaison et $\theta_{ik} = 0$ autrement. Nous définissons les poids associés standardisés $\tilde{\theta}_{ik} = \theta_{ik} \Big/ \sum_{i'\in U_F} \theta_{i'k}$ qui satisfont la contrainte $\sum_{i'\in U_F} \tilde{\theta}_{i'k} = 1$. Si pour une unité donnée $k \in U_T$, U_{Fk} désigne l'ensemble des unités i de U_F qui sont associées à k, les poids de liens associés θ_{ik} sont égaux à zéro quand i n'appartient pas à U_{Fk} et les poids standardisés associés peuvent se formuler sous la forme $\tilde{\theta}_{ik} = \theta_{ik} \Big/ \sum_{i'\in U_{Fk}} \theta_{i'k}$. Le total t_y peut s'écrire comme le total sur U_F de la variable $\sum_{k\in U_T} \tilde{\theta}_{ik} y_k, i \in U_F$ comme suit :

$$t_y = \sum_{k\in U_T} y_k = \sum_{k\in U_T} \left(\sum_{i\in U_F} \tilde{\theta}_{ik} \right) y_k = \sum_{i\in U_F} \left(\sum_{k\in U_T} \tilde{\theta}_{ik} y_k \right). \tag{3.1}$$

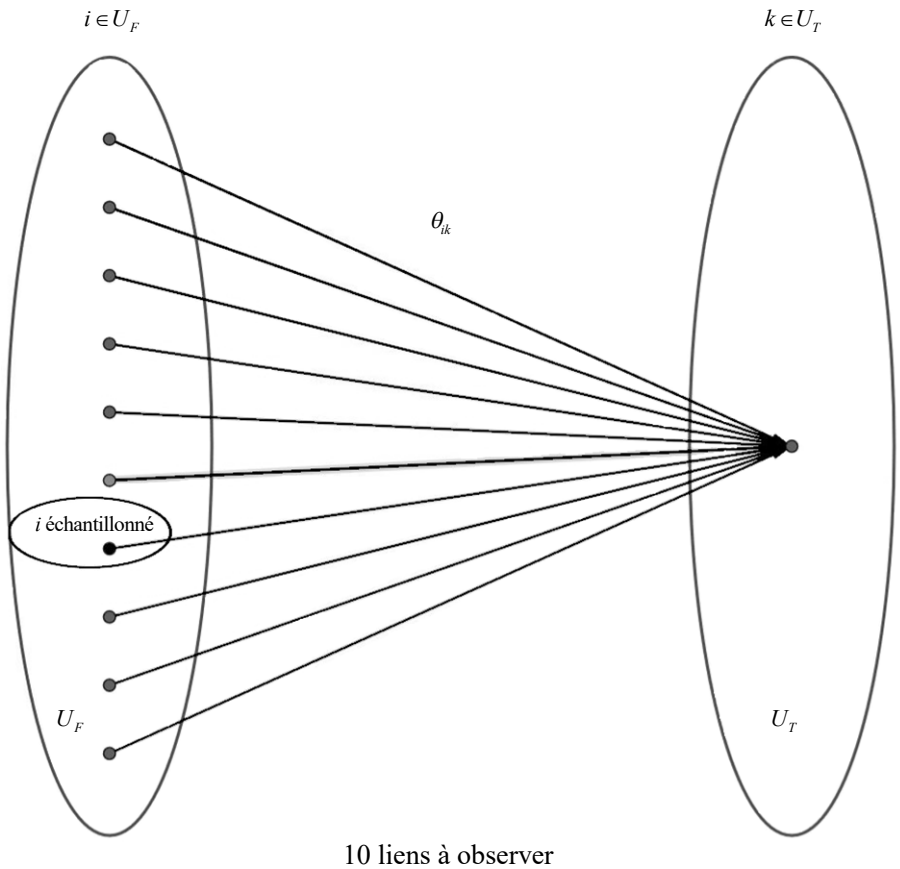
L'estimateur MGPP de t_y, étudié dans Deville et Lavallée (2006), est donné par

$$\hat{t}_{y1} = \sum_{i \in s_F} d_i \left(\sum_{k \in s_T} \tilde{\theta}_{ik} y_k \right), \tag{3.2}$$

et il estime sans biais le total t_y, à condition que les poids de liens associés soient standardisés.

Pour calculer $\hat{t}_{y1}$, les poids de liens doivent être standardisés pour tout $k \in s_T$, ce qui implique que la somme $\sum_{i' \in U_F} \theta_{i'k} = \sum_{i' \in U_{Fk}} \theta_{i'k}$ soit connue pour les unités k échantillonnées dans U_T. Dans la figure 3.1, nous considérons le cas de 10 observations dans U_F en liaison avec une seule observation dans U_T avec les poids associés θ_{ik}. L'observation i dans l'ellipse de la figure qui suit est échantillonnée dans U_F, ce qui implique que l'unité de U_T en liaison avec i est échantillonnée et que 10 liens doivent être observés pour calculer la somme des poids associés.

Figure 3.1 Petit exemple avec 10 liens « tous pour un » pour une unité échantillonnée dans la population cible à partir d'une unité échantillonnée dans la population de base.



Comme l'estimateur direct de Horvitz-Thompson, les estimateurs MGPP sont sans biais pour le total d'intérêt. Il est toutefois impossible de comparer leur variance avec un estimateur direct de Horvitz-Thompson pour un plan de sondage quelconque. C'est ce que rappellent les auteurs dans la sous-section 4.1

de Ardilly et coll. (2023) en disant que, pour l'échantillonnage indirect et avec des probabilités d'inclusion du premier ordre dans la population frame, il est habituellement impossible de calculer les probabilités d'inclusion dans la population cible.

Une comparaison entre l'échantillonnage direct et l'échantillonnage indirect est donc difficile en général. On obtient toutefois un résultat pour les liens « tous pour un » et l'échantillonnage de Poisson dans Medous et coll. (2023). Pour l'échantillonnage de Poisson dans la population frame et les liens « tous pour un », il est possible de calculer les probabilités d'inclusion dans la population cible à partir des probabilités correspondantes dans la population frame. Nous pouvons ainsi comparer la précision de l'estimateur de Horvitz-Thompson d'un total obtenue pour l'échantillonnage direct avec celle obtenue pour l'échantillonnage indirect. La proposition 2.3 dans Medous et coll. (2023) indique que la variance de l'estimateur de Horvitz-Thompson pour l'échantillonnage direct est toujours inférieure ou égale à la variance de l'estimateur MGPP pour le plan d'échantillonnage indirect associé. Pour l'échantillonnage de Poisson et les liens « tous pour un », il y a une perte de précision en utilisant l'échantillonnage indirect et un estimateur MGPP par rapport à l'échantillonnage direct et l'estimateur de Horvitz-Thompson.

Ajoutons que, dans le cas des liens « tous pour un » et pour une condition générale appelée propriété Δ sur le plan de sondage, Medous et coll. (2023) montrent que les poids de liens faiblement optimaux dans Deville et Lavallée (2006) sont aussi les poids de liens fortement optimaux pour l'estimateur MGPP (pour plus de détails, voir la section 4.5 de Ardilly et coll., 2023). La propriété Δ caractérise la matrice des covariances $\Delta = (\delta_{ii'})_{i,i' \in U_F}$ des indicatrices aléatoires d'appartenance à un échantillon pour le plan de sondage $p_F(.)$. Rappelons que, pour une unité donnée $k \in U_T$, U_{Fk} représente l'ensemble de taille N_{Fk} des unités i de U_F qui sont en liaison avec k. En raison de la structure des liens « tous pour un », la taille N_F de U_F est égale à la somme $\sum_{k \in U_T} N_{Fk}$. Dans ce qui suit, nous considérons que les unités de U_F s'ordonnent suivant les U_{Fk}, de sorte que la matrice Δ de taille $N_F \times N_F$ peut se formuler par blocs $\Delta_{kk'}$, $k, k' \in U_T$, de taille $N_{Fk} \times N_{Fk}$. Le nombre de blocs est égal à la taille N_T de la population U_T. Pour des liens « tous pour un » et des indices ordonnés dans la population U_F comme indiqué ci-dessus, un plan de sondage $p_F(.)$ satisfait la propriété Δ si :

(i) Δ_{kk} est inversible pour tout $k \in U_T$,

(ii) $\Delta_{kk'} = c_{kk'} \mathbf{1}_k \mathbf{1}_{k'}^t$ pour $k \neq k' \in U_T$, où $\mathbf{1}_k$ est le vecteur de dimension N_{Fk} contenant des valeurs 1.

La propriété Δ est une propriété technique permettant de simplifier l'expression de la variance de l'estimateur MGPP dans le cas particulier des liens « tous pour un ». Cette propriété est vraie pour l'échantillonnage de Poisson et l'échantillonnage aléatoire simple sans remise, parce que la matrice des covariances Δ de ces plans de sondage a des termes constants hors la diagonale. Mais elle est vérifiée aussi, par exemple, pour un échantillonnage aléatoire simple stratifié sans remise avec le même taux d'échantillonnage pour toutes les strates ou avec des taux différents lorsque les unités de U_F en liaison avec la même unité de U_T appartiennent à la même strate (voir les détails dans Medous et coll., 2023). Ce résultat a déjà été prouvé pour l'échantillonnage de Poisson et l'échantillonnage aléatoire simple sans remise dans

Deville et Lavallée (2006), mais n'a pas été clairement énoncé comme une propriété de la structure des liens « tous pour un ». En outre et toujours à propos de la propriété Δ, Medous et coll. (2023) offrent une expression explicite et simple pour la différence des variances entre un estimateur MGPP quelconque et l'estimateur MGPP optimal.

Le plan de sondage SYCI2 est en réalité plus complexe qu'un simple plan d'échantillonnage indirect, d'où la nécessité d'étudier plus en détail la perte de précision observée à La Poste lorsque SYCI2 remplace SYCI. SYCI2 fait appel à un plan d'échantillonnage indirect double comprenant une population intermédiaire de boîtes de tri du courrier. Les propriétés des plans d'échantillonnage indirect double ont été étudiées dans Medous et coll. (2023) et Medous (2023). Nous les expliquons brièvement dans la sous-section qui suit.

3.3 Échantillonnage indirect double « tous pour un-tous pour un »

Comme il est expliqué à la section 4 de Ardilly et coll. (2023), le total des poids de liens associés doit être connu pour chaque observation échantillonnée indirectement dans la population cible afin de standardiser les poids de liens et d'obtenir des estimateurs sans biais. Dans le cas des liens « tous pour un », le nombre de liens entre la population de base et la population cible peut être très élevé et le dénombrement de ces liens peut se révéler infaisable. À La Poste, cela signifie que toutes les adresses desservies dans une tournée de postier échantillonnée doivent être connues. On compte en moyenne quelque 500 adresses par tournée de postier, et il est impossible de dénombrer toutes ces adresses avant le départ du postier. Pour résoudre ce problème, La Poste a dû envisager un plan d'échantillonnage indirect double et un double estimateur MGPP. L'idée de cet échantillonnage indirect double est d'introduire une population intermédiaire U_M entre la population frame et la population cible et d'appliquer ensuite les mêmes principes de l'échantillonnage indirect simple.

Le recours à une population intermédiaire avait déjà été présenté dans Deville et Lavallée (2006) (voir la sous-section 3.3 pour la propriété de transitivité et la sous-section 6.1 pour la factorisation) de manière à simplifier le calcul des poids associés optimaux. À La Poste, on constitue une population intermédiaire de boîtes de tri postal pour obtenir moins de liens à observer. L'idée est qu'en introduisant une population intermédiaire, nous avons plus de liberté dans le choix des poids de liens associés, comme nous allons le préciser ci-dessous. Par souci de simplicité, nous nous attacherons au cas où non seulement les liens entre U_F et U_M, sont « tous pour un » mais aussi ceux entre U_M et U_T. Nous considérerons les poids de liens positifs $\theta_{ik}^{\mathrm{FT}}$ pour $i \in U_F$ et $k \in U_T$ (respectivement $\theta_{ij}^{\mathrm{FM}}$ pour $i \in U_F, j \in U_M$ et $\theta_{jk}^{\mathrm{MT}}$ pour $j \in U_M, k \in U_T$) associés aux liens entre U_F et U_T (respectivement U_F et U_M, et U_M et U_T). Pour exprimer le total $t_y = \sum_{k \in U_T} y_k$ comme un total sur U_F comme dans l'équation (3.1), nous devons calculer les poids de liens standardisés $\tilde{\theta}_{ij}^{\mathrm{FM}}$ et $\tilde{\theta}_{jk}^{\mathrm{MT}}$ de sorte que

$$\sum_{i \in U_F} \sum_{j \in U_M} \tilde{\theta}_{ij}^{\mathrm{FM}} \tilde{\theta}_{jk}^{\mathrm{MT}} = 1. \tag{3.3}$$

Nous avons :

$$t_y = \sum_{k\in U_T} y_k = \sum_{k\in U_T}\left(\sum_{i\in U_F}\sum_{j\in U_M}\tilde{\theta}_{ij}^{\mathrm{FM}}\tilde{\theta}_{jk}^{\mathrm{MT}}\right)y_k = \sum_{i\in U_F}\left(\sum_{k\in U_T}\sum_{j\in U_M}\tilde{\theta}_{ij}^{\mathrm{FM}}\tilde{\theta}_{jk}^{\mathrm{MT}}y_k\right).$$

Un double estimateur MGPP sans biais pour t_y est de la forme suivante :

$$\hat{t}_{y2} = \sum_{i\in s_F} d_i\left(\sum_{k\in s_T}\sum_{j\in s_M}\tilde{\theta}_{ij}^{\mathrm{FM}}\tilde{\theta}_{jk}^{\mathrm{MT}}y_k\right), \tag{3.4}$$

où l'échantillon s_M de U_M (respectivement s_T de U_T) est constitué des unités de U_M (respectivement de U_T) en liaison avec au moins une unité de s_F (respectivement de s_M).

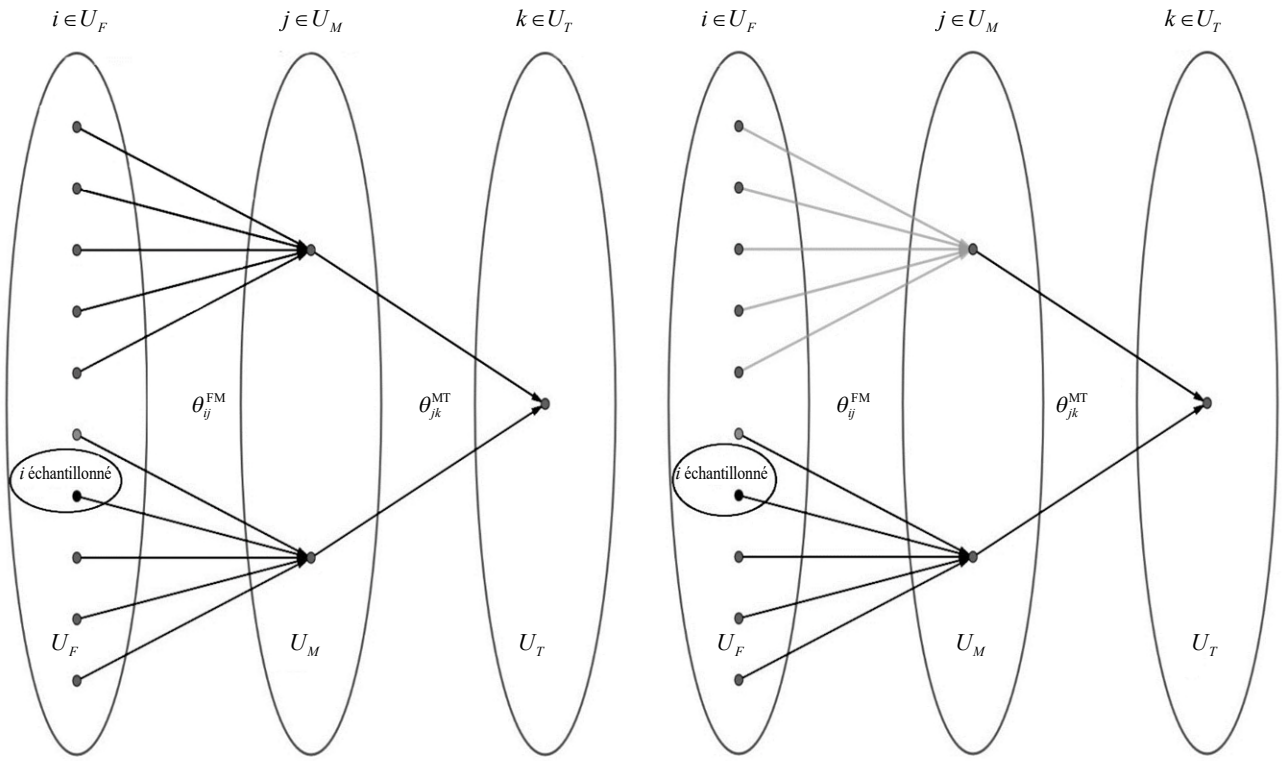
La standardisation des poids de liens associés en (3.3) est appelée « standardisation globale ». Si on standardise chaque ensemble de liens $\tilde{\theta}_{ij}^{\mathrm{FM}}$ et $\tilde{\theta}_{jk}^{\mathrm{MT}}$ séparément, appelé « double standardisation », il est facile de vérifier que les poids de liens sont aussi globalement standardisés. Mais comme l'illustre le petit exemple qui suit, avec une double standardisation des poids de liens, le nombre de liens à observer peut diminuer par rapport à un échantillonnage indirect simple. Cette caractéristique n'est pas nécessairement vraie pour d'autres types de poids de liens globalement standardisés.

Les liens entre la population de base U_F et la population cible U_T sont les mêmes à la figure 3.2 qu'à la figure 3.1, mais une population intermédiaire est introduite avec deux unités, chacune étant liée à la seule unité de U_T, et à cinq unités de U_F. Pour le calcul de l'estimateur $\hat{t}_{y2}$ en (3.4), il faut que les poids de liens associés soient globalement standardisés (voir (3.3)). Une possibilité de standardisation pour la somme $\sum_{j\in U_M}\theta_{ij}^{\mathrm{FM}}\theta_{jk}^{\mathrm{MT}}$ est de la diviser par sa somme sur $i\in U_{Fk}$, où U_{Fk} est la population d'unités de U_F associé avec $k\in U_T$. Du côté gauche à la figure 3.2, cette double standardisation conduit à observer 12 liens, puisqu'il y a 10 observations dans U_F liées à l'observation dans U_T et 2 autres liens entre U_M et U_T. Cette situation est encore pire que s'il s'agissait d'observer les 10 liens en échantillonnage indirect simple (voir la figure 3.1).

Au contraire, le recours à la double standardisation des poids de liens associées est moins coûteuse, et la raison en est quelque peu subtile. La standardisation des $\theta_{ij}^{\mathrm{FM}}$ (respectivement $\theta_{jk}^{\mathrm{MT}}$), consiste à les diviser par leur somme sur $i\in U_{Fj}$ (respectivement $j\in U_{Mk}$), où U_{Fj} (respectivement U_{Mk}) est la population d'unités de U_F (respectivement de U_M) liée à $j\in U_M$ (respectivement $k\in U_T$). En d'autres termes, les liens doivent être connus seulement pour $i\in U_{Fj}$ pour la double standardisation, par opposition à U_{Fk} pour la standardisation globale comme nous l'avons évoqué précédemment. Du côté droit de la figure 3.2, les deux liens entre U_M et U_T doivent être observés, mais seulement cinq liens (sur dix) doivent l'être entre U_F et U_M, ce qui donne sept liens à observer au total. Comme le décrivent Medous et coll. (2023), pour les liens « tous pour un » entre U_F et U_M et entre U_M et U_T, l'échantillonnage indirect double avec la double standardisation peut permettre de diminuer le nombre de liens à observer. À La Poste, on compte en moyenne 500 adresses postales par tournée de postier avec 50 adresses par casier de tri et 10 casiers par tournée. Le recours à l'échantillonnage indirect double et à la double standardisation donne en moyenne 60 liens à observer au lieu des 500 de l'échantillonnage indirect simple.

À La Poste, l'avantage de l'échantillonnage indirect double pour le nombre de liens à observer est évident, mais des questions se posent sur le choix de la double standardisation des poids de liens. Est-ce que le choix des poids de liens peut avoir une grande incidence sur l'efficacité des estimateurs MGPP ? Est-il possible de trouver des poids de liens associés doublement standardisés optimaux ? On peut répondre oui à ces deux questions. Dans Medous et coll. (2023), nous avons effectué certaines simulations de Monte Carlo pour montrer qu'il peut y avoir des situations où la perte d'efficacité est énorme. En outre, les résultats pour les données postales historiques nous permettent de dégager un facteur de deux à trois de perte d'efficacité entre SYCI etSYCI2. Très récemment, Medous (2023) a obtenu des poids de liens doublement standardisés optimaux. Ces poids de liens nous permettent d'avoir moins de liens à observer comparativement à un plan d'échantillonnage indirect simple, tout en obtenant un estimateur MGPP de variance minimale (quelle que soit la variable d'intérêt).

Figure 3.2 Même exemple qu'à la figure précédente, mais avec une population intermédiaire. À gauche, 12 liens sont nécessaires pour la standardisation globale. À droite, seuls 7 liens sont nécessaires pour la double standardisation.



Nombre de poids observés : 12 Nombre de poids observés : 7

4. Conclusion

Les deux applications que nous avons analysées sont autant d'exemples de l'héritage de Jean-Claude. Nous avons aussi travaillé avec lui sur l'estimation de paramètres complexes et sur la coordination

d'échantillons en utilisant la fonction d'influence partielle (Goga, Deville et Ruiz-Gazen, 2009). Nous nous sommes aussi inspirées de ses idées sur les méthodes non paramétriques en sondage (Goga et Ruiz-Gazen, 2014).

Jean-Claude était un homme d'échange et de transmission, avec des collaborations fructueuses en recherche et des présentations de vulgarisation fascinantes. Nous nous rappellerons en particulier de ses exposés aux étudiants du master « statistique et économétrie » à l'École d'économie de Toulouse, ainsi que de son article dans « Pour la science » (Deville, 2006). Il avait une foule d'anecdotes captivantes à raconter, sur la méthodologie rénovée du recensement français, sur l'enquête sur le tourisme en Bretagne (avec ses histoires de boulangeries et de péages sur les autoroutes), sur l'enquête d'évaluation du coût de la rénovation de la bibliothèque « François Mitterrand », pour ne citer que ces exemples.

Avec Jean-Claude disparaît une source de savoirs et d'intelligence à laquelle nous pensions pouvoir nous abreuver à l'infini, mais aussi un « père » et un ami (nous faisions partie de ses « copains » comme il disait). Il nous manque terriblement.

Remerciements

Nous remercions Alain Dessertaine pour les nombreuses discussions et les souvenirs au sujet de Jean-Claude Deville.

Bibliographie

Ardilly, P., Haziza, D., Lavallée, P. et Tillé, Y. (2023). Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle. *Techniques d'enquête*, 49, 2, 279-321. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2023002/article/00017-fra.pdf.

Besse, P.C., Cardot, H. et Stephenson, D. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27, 673-682.

Cardot, H., Chaouch, M., Goga, C. et Labruère, C. (2008). Functional principal components analysis with survey data. *Functional and Operatorial Statistics*, (Éds., S. Dabo-Niang et F. Férraty), Heidelberg: Springer-Verlag.

Cardot, H., Chaouch, M., Goga, C. et Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140, 75-91.

Cardot, H., Degras, D. et Josserand, E. (2013). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19, 2067-2097.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E. et Lardin, P. (2013). Comparaison de différents plans de sondage et construction de bandes de confiance pour l'estimation de la moyenne de données fonctionnelles : une illustration sur la consommation électrique. *Techniques d'enquête*, 39, 2, 313-331. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013002/article/11888-fra.pdf.

Cardot, H., Goga, C. et Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7, 562-596.

Cardot, H., Goga, C. et Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scandinavian Journal of Statistics*, 41, 516-534.

Cardot, H., Goga, C. et Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27, 243-260.

Cardot, H., et Josserand, E. (2011). Horvitz-Thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98, 107-118.

Chiky, R. (2009). *Résumé de Flux de Données Distribués*. PhD thesis, Paris: Sup Telecom.

Chiou, J.-M., Müller, H. et Wang, J. (2003). Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, B*, 65, 405-423.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Ann. Insee*, 15, 3-104.

Deville, J.-C. (1977). Analyse harmonique du calendrier de constitution des familles en france. Disparités sociales et évolution de 1920 à 1960. *Population (French Edition)*, 32(1), 17-63.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 2, 219-230. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4882-fra.pdf.

Deville, J.-C. (2006). Peut-on croire aux sondages ? *Pour la Science*, (344), 58-65.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 2, 185-196. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9551-fra.pdf.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Goga, C., Deville, J.-C. et Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample data. *Biometrika*, 96(3), 691-709.

Goga, C., et Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76, 113-140.

Hastie, T., et Mallows, C. (1993). A discussion on "A statistical view of some chemometrics regression tools" par L.E. Frank et J.H. Friedman. *Technometrics*, 35, 140-143.

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer Series in Statistics, second edition. New York: Springer-Verlag.

Kneip, A., et Utikal, K. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96, 519-542.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer Science & Business Media.

Medous, E. (2023). *Optimal Weights for double Many-To-One Generalized Weight Share Method*. Document préliminaire.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. et Puech, P. (2023). Many-to-One indirect sampling with application to the French postal traffic estimation. *The Annals of Applied Statistics*, 17(1), 838-859.

Ramsay, J.-O., et Silverman, B.-W. (2005). *Functional Data Analysis*, second edition. New York: Springer Series in Statistics.

Skinner, C.J., Holmes, D. et Smith, T. (1986). The effect of sample design on principal components analysis. *Journal of the American Statistical Association*, 81, 789-798.