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

Techniques d'enquête

Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle

par Pascal Ardilly, David Haziza, Pierre Lavallée et Yves Tillé

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.

Techniques d'enquête, décembre 2023
Vol. 49, nº 2, pp. 279-321
Statistique Canada, nº 12-001-X au catalogue

Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle

Pascal Ardilly, David Haziza, Pierre Lavallée et Yves Tillé[1]

Résumé

Jean-Claude Deville, décédé en octobre 2021, fut l'un des chercheurs les plus influents dans le domaine la statistique d'enquête au cours des quarante dernières années. Cet article retrace certaines de ses contributions qui ont eu un profond impact, tant sur la théorie que sur la pratique des enquêtes. Cet article abordera les sujets suivants : l'échantillonnage équilibré au moyen de la méthode du cube, le calage, la méthode du partage des poids, le développement des expressions de la variance d'estimateurs complexes au moyen de la fonction d'influence et l'échantillonnage par quotas.

Mots-clés : Calage; échantillonnage équilibré; échantillonnage par Quota; estimation de la variance; méthode du Cube; méthode du partage des poids.

1. Introduction

Jean-Claude Deville, décédé en octobre 2021, laissera sans aucun doute un legs important à la statistique d'enquête. Durant plus de 40 ans, dans le cadre de l'Insee puis de l'Ensai en France, il a enchaîné les innovations d'ampleur dont : techniques de calage, échantillonnage équilibré, échantillonnage indirect et partage des poids, calcul de variance, en particulier s'agissant d'estimateurs complexes, traitement de la non-réponse non-ignorable, enquêtes par quotas. Cela étant, il a travaillé dans tous les domaines des sondages, et même au-delà. Son exceptionnelle productivité tient essentiellement à une imagination très féconde alliée à une remarquable maîtrise de l'outil mathématique. Elle était aussi alimentée par les cas concrets qu'offrait l'Insee, qui comme tous les instituts nationaux de statistique se trouvait confronté en permanence à des contraintes diverses et à des obstacles qu'il fallait impérativement surmonter, généralement rapidement et à moindre coût. Il devait donc, en sa qualité de chef de l'unité de méthodologie statistique, relever « au fil de l'eau » les défis techniques qui lui étaient soumis.

Ce qui suit constitue un aperçu des développements originaux de Jean-Claude Deville, qui sont tous passés à la postérité et que l'on peut retrouver de manière approfondie dans les nombreux articles qu'il a publiés tout au long de sa carrière, certains étant partagés avec d'autres collègues avec qui il avait des relations privilégiées. À l'évidence, certains de ces développements ont trouvé depuis leur publication une application au niveau international d'une ampleur considérable, et on peut même parler de mise en œuvre « industrielle » pour ce qui concerne le calage, dont le développement a été conçu avec un autre statisticien prestigieux, Carl-Erik Särndal.

1. Pascal Ardilly, L'Institut national de la statistique et des études économiques (France); David Haziza, University of Ottawa (Canada). Courriel : dhaziza@uottawa.ca; Pierre Lavallée, Statistics Canada (retraité); Yves Tillé, Université de Neuchâtel (Suisse).

2. L'échantillonnage à probabilités inégales et équilibré

2.1 Innovations dans les algorithmes d'échantillonnage

Un échantillon est dit équilibré sur une variable si les estimateurs de Horvitz-Thompson des totaux calculé à partir d'un échantillon est égal ou presque égal au total de la population $U = \{1, \ldots, k \ldots, N\}$. Formellement, supposons qu'un vecteur de variables auxiliaires $\mathbf{z}_k = (z_{k1}, \ldots, z_{kQ})^\top$ soit connu sur toutes les unités de la population. Un échantillon S est équilibré sur $\mathbf{z}_k$ si

$$\sum_{k \in S} \frac{\mathbf{z}_k}{\pi_k} = \sum_{k \in U} \mathbf{z}_k,$$

où π_k est la probabilité d'inclusion, c'est-à-dire la probabilité que l'unité k soit dans l'échantillon aléatoire S.

L'idée de sélectionner un échantillon équilibré remonte au tout début de la théorie des sondages. Kiær (1896, 1899, 1903, 1905) a été le premier à proposer ce qu'il a appelé des « dénombrements représentatifs ». Il s'agit en fait d'une sélection d'échantillons par quotas. Cependant, c'est Gini et Galvani (1929) qui ont pour la première fois réalisé la sélection d'un échantillon équilibré en statistique officielle. Ceux-ci ont sélectionné 29 districts italiens (*circondari*) parmi 214 afin de restituer au mieux plusieurs moyennes de la population (Langel et Tillé, 2011; Tillé, 2016; Brewer, 2013). Cette méthode a été durement critiquée par Jerzy Neyman parce que l'échantillon n'était pas choisi au hasard (Bellhouse, 1988, voir). Yates (1949) and Thionet (1953) ont proposé des méthodes pour lesquelles un échantillon est sélectionné puis amélioré en remplaçant successivement des unités pour s'approcher d'une situation d'équilibrage. Hájek (1964, 1981) a proposé d'utiliser le tirage réjectif qui consiste à sélectionner une suite d'échantillons jusqu'à l'obtention d'un échantillon suffisamment équilibré. Cette méthode a cependant l'inconvénient de modifier les probabilités d'inclusion des unités sans qu'il soit possible de les calculer ensuite avec exactitude (Choudhry et Singh, 1979; Dupačová, 1979; Fuller, 2009; Legg et Yu, 2010; Boistard, Lopuhaä et Ruiz-Gazen, 2012; Fuller, Legg et Li, 2017).

Jean-Claude Deville s'est rapidement intéressé aux méthodes d'échantillonnage. En 1987, il publie un chapitre de livre avec Jean-Marie Grosbras où les méthodes d'échantillonnage sont décrites et comparées (Deville et Grosbras, 1987). L'année suivante, ceux-ci proposent avec Nicole Roth une première méthode d'échantillonnage équilibré (Deville, Grosbras et Roth, 1988). La méthode s'applique seulement au tirage à probabilités égales. L'idée consiste à découper l'espace des variables en quadrants et à sélectionner à chaque étape une unité dans le quadrant qui contribuera le plus à respecter l'équilibrage. Dans les actes de la conférence d'Örebro qui s'est tenue à l'Institut de Statistique de Suède en 1992, Jean-Claude Deville fait déjà part de son point de vue sur les trois facettes de l'utilisation d'informations auxiliaires que sont les échantillons contraints (c'est-à-dire équilibrés), l'inférence conditionnelle et la pondération (Deville, 1992).

Parallèlement à ces travaux, Jean-Claude Deville a mené une recherche très pointue sur les questions d'échantillonnage. Il propose une formalisation de l'échantillonnage dans une population continue (Deville,

1989) bien avant la publication de Cordy (1993) qui est souvent cité comme première référence de ce domaine. Il propose aussi une méthode de tirage à probabilités inégales (Deville, 1998c) qui est une variante de l'échantillonnage systématique.

Ensuite, avec Yves Tillé, il propose la méthode de scission (Deville et Tillé, 1998) pour sélectionner des échantillons à probabilité inégales. Cette famille de méthodes consiste à prendre le vecteur $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$ de probabilités d'inclusion dont la somme est égale à un entier n. Puis, à chaque étape $t, 0, 1, 2, \ldots,$ ce vecteur $\boldsymbol{\pi}(t)$ est modifié aléatoirement jusqu'à l'obtention d'un vecteur ne contenant que des valeurs égales à 0 ou à 1, ce qui correspond à la sélection d'un échantillon. Pour que la méthode soit correcte, il faut que trois conditions soient respectées :

1. Toutes les composantes des $\boldsymbol{\pi}(t)$ restent dans l'intervalle [0, 1].
2. La somme des composantes des $\boldsymbol{\pi}(t)$ reste égale à n.
3. La propriété de martingale doit être satisfaite

$$\mathrm{E}_p\{\boldsymbol{\pi}(t) \mid \boldsymbol{\pi}(t-1)\} = \boldsymbol{\pi}(t-1), \quad \text{pour tout } t, \tag{2.1}$$

où $\mathrm{E}_p(\cdot)$ est l'espérance mathématique sous le plan de sondage qui prend en compte l'aléa dû à l'échantillonnage.

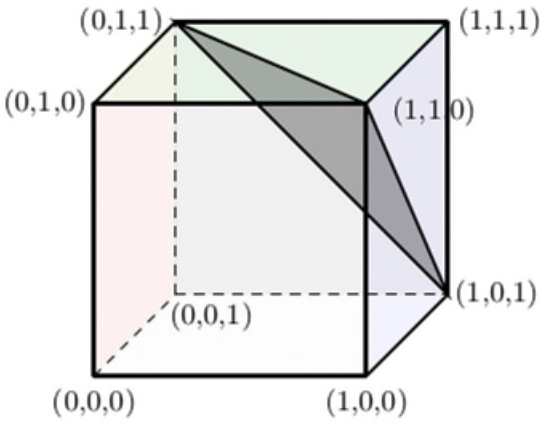
La propriété de martingale suffit pour montrer que les probabilités d'inclusion sont respectées à chaque étape. En effet, par le théorème de l'espérance totale, on a directement $\mathrm{E}_p\{\boldsymbol{\pi}(t)\} = \boldsymbol{\pi}(0)$.

La méthode de scission est une manière très générale de représenter une méthode d'échantillonnage. Presque tous les algorithmes de tirage peuvent être réécrits sous forme de scission. Cette méthode permet de se concentrer sur une étape élémentaire. Le contrôle des trois conditions permet de vérifier rapidement si la méthode est correcte ou non.

Une des méthodes proposées comme cas particulier de la méthode de scission est la méthode du pivot pour laquelle seulement deux composantes des vecteurs $\boldsymbol{\pi}(t)$ sont modifiées à chaque étape. Cette méthode a été généralisée pour sélectionner plusieurs échantillons qui ne se chevauchent pas dans la même population avec des probabilités égales ou inégales (Deville et Tillé, 2000b).

Le passage de la méthode de scission à l'échantillonnage équilibré a été relativement simple quand Jean-Claude Deville et Yves Tillé se sont aperçus que les échantillons $(I_1, \ldots, I_N)^\top$ codés sous forme de vecteurs contenant uniquement des 0 et des 1 sont les sommets d'un N-cube de $\mathbb{R}^N$. De plus, les conditions 1 et 2 de la méthode de scission peuvent être interprétées géométriquement. Les vecteurs $\boldsymbol{\pi}(t)$ doivent rester dans le simplexe $\mathcal{P} = \left\{ c_k \in [0,1] \,\middle|\, \sum_{k=1}^{N} c_k = n \right\}$. La figure 2.1 contient une représentation de ce simplexe pour un échantillonnage de taille $n = 2$ dans une population de taille $N = 3$. La méthode de scission est donc une promenade aléatoire dans un simplexe qui doit respecter la propriété de martingale.

Figure 2.1 Simplexe réunissant les échantillons de taille $n = 2$ dans une population de taille $N = 3$ à l'intérieur d'un cube dont les échantillons sont les sommets. Le simplexe est ici un triangle équilatéral.



2.2 L'échantillonnage équilibré par la méthode du cube

Le passage à l'échantillonnage équilibré est alors apparu comme une évidence. Il suffit de remplacer la condition 2 de la méthode de scission pour obtenir les principes généraux d'une méthode d'échantillonnage équilibré. On obtient dès lors les trois conditions suivantes :

1. Toutes les composantes des $\boldsymbol{\pi}(t)$ restent dans l'intervalle [0,1].
2. À chaque étape $t = 0, 1, 2, \ldots$ les vecteurs $\boldsymbol{\pi}(t) = (\pi_1(t), \ldots, \pi_N(t))^\top$ doivent satisfaire aux équations d'équilibrage :

$$\sum_{k \in U} \frac{\mathbf{z}_k}{\pi_k} \pi_k(t) \;=\; \sum_{k \in U} \mathbf{z}_k .$$

3. La propriété de martingale doit être satisfaite

$$\mathrm{E}_p\{\boldsymbol{\pi}(t) \,|\, \boldsymbol{\pi}(t-1)\} \;=\; \boldsymbol{\pi}(t-1), \quad \text{pour tout } \; t. \tag{2.2}$$

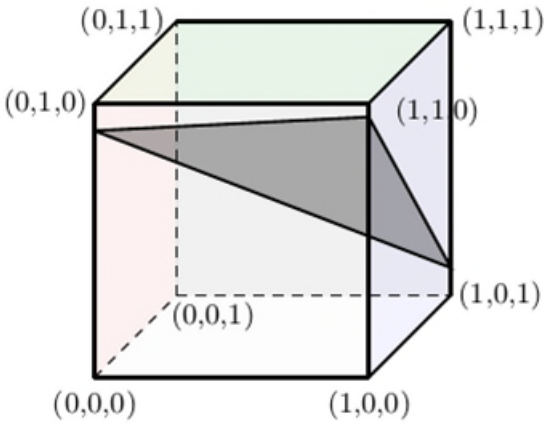
Les conditions 1 et 2 définissent maintenant le polytope

$$\mathcal{P} \;=\; \left\{ c_k \in [0,1] \,\middle|\, \sum_{k \in U} \frac{\mathbf{z}_k}{\pi_k} c_k = \sum_{k \in U} \mathbf{z}_k \right\},$$

dans lequel les $\boldsymbol{\pi}(t)$ doivent rester afin de vérifier à chaque étape les contraintes d'équilibrage. Un exemple de polytope est présenté dans la figure 2.2. Cependant, quand les contraintes sont complexes, les sommets du polytope $\mathcal{P}$ ne sont pas nécessairement des sommets du cube, ce qui signifie qu'il peut ne pas exister d'échantillons exactement équilibrés. On ne pourra dès lors qu'obtenir un échantillon approximativement équilibré. C'est pourquoi la méthode du cube est composée de deux phases : la phase de vol et la phase d'atterrissage.

La phase de vol est une promenade aléatoire dans le polytope $\mathcal{P}$ qui se termine sur un des sommets du polytope. La phase d'atterrissage consiste à choisir un échantillon approximativement équilibré proche du sommet du polytope obtenu à la fin de la phase de vol tout en respectant les probabilités d'inclusion.

Figure 2.2 Polytope $\mathcal{P}$ pour le cas où les sommets du polytope ne sont pas un sommet du cube.



La méthode du cube est une famille de méthode qui permet de générer une telle promenade aléatoire. Pour la phase de vol, afin de passer de $\boldsymbol{\pi}(t)$ à $\boldsymbol{\pi}(t+1)$, la méthode du cube procède de la manière suivante.

1. On génère un vecteur $\mathbf{u}(t) = (u_1(t), \ldots, u_N(t))^\top$ tel que

$$\sum_{k \in U} \frac{\mathbf{z}_k}{\pi_k} u_1(t) = \mathbf{0},$$

et $u_k(t) = 0$, si $\pi_k(t)$ est un nombre entier (0 ou 1). Si un tel vecteur n'existe pas, la phase de vol s'arrête.

2. On cherche λ_1 et λ_2 les plus grandes valeurs positives qui satisfont

$$0 \le \pi_k(t) + \lambda_1 u_k(t) \le 1 \quad \text{et} \quad 0 \le \pi_k(t) - \lambda_2 u_k(t) \le 1, \quad \text{pour tout} \quad k \in U.$$

3. On met à jour

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1 \mathbf{u}(t) & \text{avec une probabilité } q \\ \boldsymbol{\pi}(t) - \lambda_2 \mathbf{u}(t) & \text{avec une probabilité } 1-q, \end{cases}$$

où $q = \lambda_2 / (\lambda_1 + \lambda_2)$.

Il existe plusieurs manières de générer le vecteur $\mathbf{u}(t)$, ce qui permet de définir plusieurs variantes de la méthode. Après N étapes au plus, la phase de vol s'arrête sur un sommet du polytope $\mathcal{P}$. Ce sommet est

un vecteur contenant au plus Q valeurs qui ne valent ni 0 ni 1, où Q est le nombre de variables auxiliaires. Pour obtenir un échantillon, il faut alors appliquer une phase d'atterrissage. Deux variantes sont proposées dans Deville et Tillé (2004).

La méthode du cube a été publiée pour la première fois dans les actes des Journées de méthodologie statistique (Deville et Tillé, 2000a) puis comme chapitre d'un livre (Deville et Tillé, 2001). La publication en anglais a été beaucoup plus laborieuse mais a finalement été acceptée dans *Biometrika* (Deville et Tillé, 2004, 2005). Un(e) arbitre ne pouvait pas admettre que les échantillons pouvaient être en même temps équilibrés et aléatoires. Un(e) autre ne pouvait pas admettre que la méthode fonctionnait sans énumérer tous les échantillons possibles. Une autre critique était que la méthode ne fournissait pas d'échantillons exactement équilibrés. L'existence d'une solution exacte ne dépend pourtant pas de la méthode mais de la géométrie du problème.

2.3 Implémentation, applications de la méthode et prolongements de la recherche

Un premier prototype d'une fonction SAS-IML a été écrite par trois étudiants de l'*École nationale de la statistique et de l'analyse de l'information* (ENSAI) (Bousabaa, Lieber et Sirolli, 1999) sous la direction de Frédéric Tardieu et Yves Tillé. La toute première version était vraiment lente au point de douter qu'elle soit applicable, mais rapidement des progrès ont été réalisés. Chauvet et Tillé (2006a) ont proposé une implémentation qui ne considère à chaque étape qu'une petite partie de la population, ce qui réduit considérablement le temps de calcul. Une fonction SAS a été écrite en utilisant cette procédure (Chauvet et Tillé, 2006b). Plusieurs packages R permettent aussi de sélectionner directement un échantillon équilibré (Tillé et Matei, 2021; Grafström et Lisic, 2019; Jauslin, Eustache, Panahbehagh et Tillé (2021). Leur utilisation est particulièrement simple, car les fonctions ne dépendent que de deux arguments : la matrice des variables d'équilibrage et le vecteur de probabilités d'inclusion.

Jean-Claude Deville a joué un rôle décisif pour changer la procédure du recensement en France afin de passer à un système continu (Deville et Jacod, 1996). La méthode du cube a été un outil précieux pour construire les groupes de rotation (Durr et Dumais, 2002). Les unités primaires de l'échantillon maître ont également été sélectionnés avec la méthode du cube. La méthode a été très rapidement utilisée dans de nombreuses applications (Tillé, 2011). Comme pour le calage, l'équilibrage est devenu une procédure standard en statistique d'enquête.

La méthode du cube a également suscité de nombreux travaux académiques. La précision de l'équilibrage est examinée dans Chauvet, Haziza et Lesage (2015). Leuenberger, Eustache, Jauslin et Tillé (2022) proposent de trier les observations par ordre croissant de profondeur dans le nuage de point, ce qui réduit le problème d'arrondi. La question des probabilités d'inclusion optimales est traitée dans Nedyalkova et Tillé (2008, 2012) et Chauvet, Bonnéry et Deville (2011). Ces résultats généralisent la stratification optimale de Neyman (1934). Plusieurs articles traitent de l'équilibrage pour des populations stratifiées (Chauvet, 2009; Hasler et Tillé, 2014; Jauslin, Eustache et Tillé, 2021).

Plusieurs travaux ont été dédiés à l'échantillonnage spatial. Grafström, Lundström et Schelin (2012) utilisent le caractère répulsif de la méthode du pivot pour obtenir des échantillons bien étalés dans l'espace, ce qui augmente la précision quand les données sont autocorrélées. Grafström et Tillé (2013) proposent ensuite une variante de la méthode du cube afin d'obtenir des échantillons à la fois bien étalés dans l'espace et équilibrés sur des totaux. Enfin, Jauslin et Tillé (2020a, b), équilibrent sur des micro-strates contenant le voisinage de chaque unité afin d'obtenir des échantillons particulièrement bien étalés.

Jean-Claude Deville a beaucoup travaillé sur les plans à entropie maximale, question sur laquelle il a laissé plusieurs notes manuscrites (Deville, 2000b; Deville, nda, ndb, ndc). Ces résultats ont enfin permis une implémentation relativement rapide de ce plan. Deville et Qualité (2005) ont ensuite proposé une généralisation au cas multidimensionnel. Suite à la remarque d'un arbitre lors de la soumission de l'article sur la méthode du cube, Jean-Claude Deville s'est attelé à la détermination d'une condition nécessaire et suffisante pour que l'équilibrage n'ait pas de problème d'arrondi (Deville, 2015, 2014). La condition obtenue est malheureusement très restrictive. Dans le cas où la condition est respectée, il développe les plans équilibrés à entropie maximale (Deville, 2014).

Jean-Claude Deville a rapidement compris l'intérêt de la méthode du cube pour d'autres applications que l'échantillonnage. Plusieurs méthodes d'imputation équilibrées ont été proposées par Chauvet, Deville et Haziza (2011); Hasler et Tillé (2016); Eustache, Vallée et Tillé (2022). Ces méthodes ont l'avantage de bien restituer la distribution de la variable imputée tout en réduisant la variance due à l'imputation aléatoire. La méthode du cube est utilisée aussi dans des champs d'application très éloignée de l'échantillonnage comme par exemple dans les méthodes MCMC (Markov chain Monte Carlo) (Chopin et Ducrocq, 2021).

3. Le calage

Les articles écrits par Deville et Särndal (1992) et Deville, Särndal et Sautory (1993) portant sur les méthodes de calage (aussi appelées méthodes de redressement) et publiées dans la prestigieuse revue *Journal of the American Statistical Association* sont considérés comme deux des articles les plus importants et influents des trente dernières années dans le domaine de l'échantillonnage et de la statistique officielle. Ces deux articles proposent une théorie unifiée de l'estimation en présence d'information auxiliaire dont les prémisses sont discutées dans Lemel (1976) et Huang et Fuller (1978). Les deux articles co-écrits par Jean-Claude Deville ont généré de nombreux articles de recherche au cours des trois dernières décennies. Le lecteur est renvoyé à Särndal (2007), Haziza et Beaumont (2017), Devaud et Tillé (2019) et Zhang, Han et Wu (2022) pour des revues sur les méthodes de calage. La post-stratification (e.g., Holt et Smith, 1979), les méthodes de ratissage ou *raking ratio* en anglais (Deming et Stephan, 1940; Stephan, 1942), l'estimation par la régression généralisée (voir, par exemple, Särndal, Swensson et Wretman, 1992) peuvent être obtenues comme des cas particuliers des méthodes de calage.

Les méthodes de calage mettent à profit une information auxiliaire disponible à l'étape de l'estimation dans le but de garantir la cohérence entre les estimations produites par l'enquête et des totaux externes

connus ou estimés. En pratique, les méthodes de calage sont également utilisées afin de réduire les erreurs de non-réponse et les erreurs de couverture.

3.1 Le calage en l'absence d'erreurs non dues à l'échantillonnage

Dans cette section, on se placera dans un cadre théorique idéal, pour lequel les erreurs de non-réponse et les erreurs de couvertures sont supposées négligeables. Le calage repose sur la disponibilité d'un vecteur de variables auxiliaires, $\mathbf{x}_k = (x_{1k},\ldots,x_{Jk})^\top$, et du vecteur des totaux de population correspondant, $\mathbf{t}_\mathbf{x} = (t_{x_1},\ldots,t_{x_J})^\top$, où $t_{x_j} = \sum_{k\in U} x_{jk},\ j = 1,\ldots,J.$ Le vecteur $\mathbf{t}_\mathbf{x}$ est obtenu à partir d'une source externe telle que le recensement, un fichier administratif ou encore une autre enquête.

Lorsque l'on sélectionne un échantillon S d'une population $U,$ il est pratiquement certain que ce dernier souffrira d'une distorsion aléatoire en termes du vecteur de variables auxiliaires $\mathbf{x},$ au sens où $\hat{\mathbf{t}}_{\mathbf{x},\pi} \neq \mathbf{t}_\mathbf{x},$ avec $\hat{\mathbf{t}}_{\mathbf{x},\pi} = \sum_{k\in S} \mathbf{x}_k/\pi_k$. Contrairement à une distorsion systématique (comme celle que l'on observe généralement dans un contexte de non-réponse), on est ici dans un cas de distorsion aléatoire car $\mathrm{E}_p(\hat{\mathbf{t}}_{\mathbf{x},\pi}) - \mathbf{t}_\mathbf{x} = \mathbf{0}.$ Le but du calage est donc de corriger cette distorsion.

Plus formellement, on recherche un ensemble de poids de calage $\{w_k; k \in S\}$ tel que

$$\sum_{k\in S} \frac{d_k G(w_k/d_k)}{q_k} \tag{3.1}$$

est minimum sous les J contraintes de calage

$$\sum_{k\in S} w_k \mathbf{x}_k = \mathbf{t}_\mathbf{x}, \tag{3.2}$$

où $d_k = 1/\pi_k$ et q_k est un facteur d'échelle choisi par l'utilisateur (voir Deville et Särndal, 1992; Deville et coll., 1993). Dans la majorité des cas rencontrés en pratique, on pose $q_k = 1$ pour tout k. La fonction $G(\cdot)$ est une fonction de pseudo-distance permettant de mesurer la proximité entre les poids avant calage d_k et les poids après calage w_k.

Les poids de calage w_k sont donnés par

$$w_k = d_k F(q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k), \tag{3.3}$$

où $\hat{\boldsymbol{\lambda}}$ est un vecteur de taille J de coefficients estimés garantissant que les contraintes (3.2) sont satisfaites, et $F(\cdot) = g^{-1}(\cdot)$ est la fonction de calage, définie comme la fonction inverse de $g(.) \equiv \partial G(t)/\partial t$. Le poids calé (3.3) peut être vu comme le produit du poids avant calage, $d_k,$ et un facteur d'ajustement, $F(q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$. En plus du vecteur $\hat{\boldsymbol{\lambda}},$ ce dernier dépend de la fonction de calage $F(\cdot)$ (et donc de la fonction de pseudo-distance $G(\cdot)$) ainsi que des caractéristiques de l'unité k, q_k et $\mathbf{x}_k$. Dans certaines situation, le terme $q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k$ ne dépend pas de $k,$ auquel cas toutes les fonctions de calage conduiront au même ensemble de poids w_k. Cette situation survient dans le cas d'une post-stratification ou de l'estimation par le ratio (Haziza et Beaumont, 2017).

L'estimateur par calage de t_y est donné par

$$\hat{t}_{y,C} = \sum_{k \in S} w_k y_k. \tag{3.4}$$

Deville et Särndal (1992) ont considéré un éventail de fonctions $G(\cdot)$ dont certaines sont présentées au Tableau 3.1. Dans le cas de la distance du chi-deux généralisée, Deville et Särndal (1992) ont montré que les poids de calage sont donnés par

$$w_k = d_k(1 + q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k),$$

où

$$\hat{\boldsymbol{\lambda}} = \left(\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k^\top \right)^{-1} (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x},\pi}).$$

Il en découle que l'estimateur par calage coïncide avec l'estimateur bien connu par la régression linéaire généralisée (voir, par exemple, Särndal et coll., 1992)

$$\hat{t}_{y,C} = \hat{t}_{y,\pi} + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x},\pi})^\top \hat{\mathbf{B}}, \tag{3.5}$$

où

$$\hat{\mathbf{B}} = \left(\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} d_k \mathbf{x}_k q_k y_k.$$

Ce résultat constitue l'une des avancées importantes dans le domaine de l'estimation en présence d'information auxiliaire : il est possible de construire l'estimateur par la régression généralisée au moyen d'un calage. Deville et coll. (1993) ont établi que l'utilisation de l'information de Kullback-Leibler (voir Tableau 3.1) conduit à l'estimateur par ratissage classique, ce qui constitue un autre résultat majeur. Les distances du chi-deux généralisée tronquée et du logit (voir Tableau 3.1) permettent d'imposer des bornes sur les ajustements de calage, $F(q_k \hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$, dans le but de limiter la dispersion des poids.

Bien que les estimateurs de calage soient biaisés par rapport au plan de sondage, ils sont convergents, ce qui est une propriété désirable (Deville et Särndal, 1992). Lorsque la taille de l'échantillon n est suffisamment grande, le biais quadratique des estimateurs de calage devient négligeable devant leur variance. Par conséquent l'erreur quadratique moyenne des estimateurs de calage est approximativement égale à leur variance, pourvu que n soit suffisamment grand.

Deville et Särndal (1992) ont montré que la variance d'un estimateur de calage peut être approchée par

$$V_p(\hat{t}_{y,C}) \approx \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{E_k}{\pi_k} \frac{E_\ell}{\pi_\ell}, \tag{3.6}$$

où $E_k = y_k - \mathbf{x}_k^\top \mathbf{B}$ est le « résidu de population » associé à l'unité k avec

$$\mathbf{B} = \left(\sum_{k \in U} \mathbf{x}_k q_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U} \mathbf{x}_k q_k y_k .$$

Il s'agit d'une propriété remarquable : tous les estimateurs par calage ont la même variance asymptotique quelle que soit la fonction de calage $F(\cdot)$. L'expression (3.6) suggère que les estimateurs de calage sont efficaces lorsque les résidus E_k sont petits, ce qui surviendra lorsque la relation entre la variable d'intérêt y et les variables de calage $\mathbf{x}$ est linéaire et forte. Qu'en est-il si la relation n'est pas linéaire ? Dans ce cas, il est possible que le modèle n'ajuste pas bien les données, conduisant alors à de grands résidus et une grande variance. Ceci a amené Wu et Sitter (2001) à proposer une procédure calage assistée d'un modèle et qui permet de prendre en compte des relations non-linéaires au moyen, par exemple, de modèles linéaires généralisés. Cependant, contrairement au calage classique de Deville et Särndal (1992), le calage assisté d'un modèle requiert la disponibilité du vecteur $\mathbf{x}$ pour toutes les unités de la population. Cette exigence n'est généralement pas satisfaite en pratique, en particulier pour les enquêtes auprès des ménages.

Tableau 3.1
Quelques fonctions de distance introduites dans Deville et Särndal (1992).

	Fonction de distance $G(w_k/d_k)$	**Facteur d'ajustement de calage $F(q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$**
Distance du chi-deux généralisée	$\frac{1}{2}\left(\frac{w_k}{d_k}-1\right)^2$	$1+q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k$
Information de Kullback-Leibler	$\frac{w_k}{d_k}\log\frac{w_k}{d_k}-\frac{w_k}{d_k}+1$	$\exp(q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k)$
Information de Kullback-Leibler inverse	$\log\frac{d_k}{w_k}+\frac{w_k}{d_k}-1$	$\frac{1}{1-q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k}$
Distance de Hellinger	$2\left\{\sqrt{\frac{w_k}{d_k}}-1\right\}^2$	$\frac{1}{\sqrt{1-2q_k\hat{\boldsymbol{\lambda}}^\top \mathbf{x}_k}}$
Distance du chi-deux généralisée tronquée	$\begin{cases}\frac{1}{2}\left(\frac{w_k}{d_k}-1\right)^2 & L<\frac{w_k}{d_k}<M \\ \infty & \text{sinon}\end{cases}$	$\begin{cases}1+q_k\boldsymbol{\lambda}^\top \mathbf{x}_k & (L-1)\le q_k\boldsymbol{\lambda}^\top \mathbf{x}_k\le(M-1) \\ M & q_k\boldsymbol{\lambda}^\top \mathbf{x}_k>(M-1) \\ L & q_k\boldsymbol{\lambda}^\top \mathbf{x}_k<(L-1)\end{cases}$
Distance du logit	$\begin{cases}\left(a_k\log\frac{a_k}{1-L}+b_k\log\frac{b_k}{M-1}\right)\frac{d_k}{A} & L<\frac{w_k}{d_k}<M \\ \infty & \text{sinon,}\end{cases}$	$\frac{L(M-1)+M(1-L)\exp(Aq_k\boldsymbol{\lambda}^\top \mathbf{x}_k)}{M-1+(1-L)\exp(Aq_k\boldsymbol{\lambda}^\top \mathbf{x}_k)}$

Dans les enquêtes à plusieurs degrés ou plusieurs phases, on dispose de plusieurs niveaux d'information auxiliaire. Par exemple, dans le cas d'une enquête à deux degrés, l'information auxiliaire pourrait être constituée d'information au niveau des ménages (nombre d'individus dans le ménage, le nombre d'individus dans chaque groupe d'âge, le statut propriétaire/locataire, etc.) et d'information au niveau des individus

(genre, groupe d'âge, etc.). Le lecteur est renvoyé à Sautory et Le Guennec (2003) et Estevao et Särndal (2002, 2006) pour un traitement des méthodes de calage dans un contexte d'enquêtes à plusieurs degrés/phases.

3.2 Correction de la non-réponse par calage

La post-stratification et les méthodes de ratissage ont longtemps été utilisées afin de traiter la non-réponse totale; voir, par exemple, Thomsen (1978), Bethlehem et Keller (1987) et Bethlehem (1988). Les premiers travaux portant sur une approche unifiée par calage en présence de non-réponse sont présentées dans Deville et Dupont (1993) et Dupont (1993). L'approche a été plus amplement étudiée par Lundström et Särndal (1999) et Särndal et Lundström (2005). L'idée est d'obtenir des poids finaux w_k à partir des poids initiaux d_k de manière à atteindre les deux objectifs suivants : (i) réduire le biais de non-réponse et (ii) garantir la cohérence entre les estimations issues de l'enquête et des totaux connus au niveau de la population.

On considère une population U dans laquelle un échantillon S est sélectionné dont seulement les unités du sous-ensemble S_r ont répondu. On a donc $S_r \subset S \subset U$. Nous distinguons deux niveaux d'information auxiliaire : (1) le vecteur $\mathbf{x}_{Uk}$ qui est observé pour $k \in S_r$ et dont le vecteur des totaux au niveau de la population, $\sum_{k\in U} \mathbf{x}_{Uk}$, est connu. (2) le vecteur $\mathbf{x}_{Sk}$ qui est observé pour $k \in S$ et dont le vecteur des estimations de type Horvitz-Thompson, $\sum_{k\in S} d_k \mathbf{x}_{Sk}$, est disponible. Les variables $\mathbf{x}_{Uk}$ sont celles qui permettront d'assurer la cohérence entres les estimations issues de l'enquête et les totaux connus au niveau de la population. Idéalement, les variables $\mathbf{x}_{Sk}$ sont celles qui expliquent à la fois le statut de réponse R_k et les variables d'intérêt. Pour chaque $k \in S_r$, nous créons le vecteur empilé $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{Uk} \\ \mathbf{x}_{Sk} \end{pmatrix}$. On recherche un système de pondération final, $\{w_k; k \in S_r\}$, tel que

$$\sum_{k\in S_r} \frac{d_k G(w_k/d_k)}{q_k}$$

est minimum sous les contraintes de calage

$$\sum_{k\in S_r} w_k \mathbf{x}_k = \begin{pmatrix} \sum_{k\in U} \mathbf{x}_{Uk} \\ \sum_{k\in S} d_k \mathbf{x}_{Sk} \end{pmatrix}.$$

Les poids finaux sont donnés par

$$w_k = d_k \times F(q_k \hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k).$$

L'estimateur de t_y est donné par

$$\hat{t}_{y,C} = \sum_{k\in S_r} \left\{ d_k \times F(q_k \hat{\boldsymbol{\lambda}}_r^\top \mathbf{x}_k) \right\} y_k. \tag{3.7}$$

Bien que n'importe quelle fonction de pseudo-distance $G(\cdot)$ puisse être utilisée, une mise en garde est de mise. En effet, choisir une fonction de pseudo-distance dans un contexte de non-réponse revient à imposer un modèle paramétrique décrivant la relation entre l'inverse des probabilités de réponse et le vecteur $\mathbf{x}$ (Haziza et Lesage, 2016). En général, un choix erroné de la fonction $G(\cdot)$ conduira généralement à un estimateur par calage biaisé. Une exception à cette règle surviendra lorsque la variable d'intérêt y est linéairement liée au vecteur $\mathbf{x}$ et que la non-réponse est du type Missing At Random (Rubin, 1976).

Une autre contribution importante de Jean-Claude Deville est le calage sur variables instrumentales, également appelé calage généralisé (Deville, 1998a, 2000a, 2002). Cette approche a également été étudiée et discutée, entre autres, par Sautory et Le Guennec (2003), Kott (2006), Chang et Kott (2008), Kott et Chang (2010), Haziza et Beaumont (2017) et Lesage, Haziza et D'Haultfoeuille (2019). Cette approche est particulièrement utile dans un contexte de non-réponse non-ignorable (Rubin, 1976). Dans ce cas, la probabilité de réponse dépend de variables complètement observées mais également de variables disponibles pour les répondants seulement. Il s'ensuit que l'estimation des probabilités de réponse n'est pas aisée. Le calage généralisé conduit à un estimateur convergent d'un total si les conditions appelées *exclusion restriction conditions* en anglais sont satisfaites.

4. La méthode du partage des poids

Le sondage indirect consiste à sélectionner un échantillon d'une population cible à partir d'une base de sondage différente, mais reliée de quelque sorte à cette population cible. Bien des développements reliés au sondage indirect se trouvent dans les livres de Lavallée (2002, 2007) auxquels s'ajoutent des contributions plus récentes telles Deville et Maumy-Bertrand (2006), Falorsi, Piersante et Bako (2016), Kiesl (2010), Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine et Puech (2023). On verra que Jean-Claude Deville a joué un rôle de premier plan dans le développement du sondage indirect.

4.1 Les tout débuts : les enquêtes longitudinales

La genèse du sondage indirect se rapporte à un problème de pondération relié aux enquêtes longitudinales. Il s'agissait alors de pondérer les individus interrogés dans le cadre d'une enquête longitudinale sociale où l'on suit des individus appartenant à des ménages au cours du temps.

Après une sélection de ménages (et donc d'individus) à la première vague, l'évolution de la composition des ménages au fil des vagues en raison notamment des mariages et des décès survenus compliquait la pondération de l'enquête. La solution s'obtient par l'utilisation de la méthode du partage des poids (voir Lavallée, 1995).

Le problème de pondération des enquêtes longitudinales auprès de ménages a suscité l'intérêt de plusieurs auteurs, notamment, par Huang (1984), Judkins, Hubble, Dorsch, McMillen et Ernst (1984), Ernst,

Hubble et Judkins (1984), Ernst (1989) et Kalton et Brick (1995). L'article de Ernst (1989) a bien décrit la base du problème et a proposé la solution reliée au partage des poids.

Supposons une enquête longitudinale d'individus tirés auprès de ménages. On dispose des deux vagues de données : la vague A (ou première vague) et la vague B (une vague subséquente). Un échantillon S^A contenant m^A individus a été tiré de la population de la vague A contenant M^A individus. Soit $\pi_k^A > 0$, la probabilité de sélection de l'individu k. À la vague B, la population contient alors M^B individus répartis dans N^B ménages U_i^B, où le ménage i contient M_i^B individus.

Le processus de l'enquête longitudinale est le suivant. Pour chaque individu k de S^A, on établit la liste des M_i^B individus du ménage i de la vague B contenant cet individu. Soit S^B, l'ensemble des n^B ménages identifiés par les individus $k \in S^A$. Une fois les ménages de S^B identifiés, on enquête auprès de tous les individus k des ménages $i \in S^B$ pour mesurer la variable d'intérêt y. La méthode du partage des poids attribue un poids d'estimation w_{ik} à chaque individu k d'un ménage enquêté U_i^B. Les étapes de la méthode sont les suivantes :

- **Étape 1** Pour chaque individu k des ménages i de S^B, on calcule le poids initial $w'_{ik} = \gamma_k / \pi_k^A$, où $\gamma_k = 1$ si $k \in S^A$, et 0 sinon.
- **Étape 2** Pour chaque ménage i de S^B, on obtient le nombre total d'individus M_i^{AB} du ménage i présents à la vague A (mais pas nécessairement contenu dans S^A).
- **Étape 3** On calcule le poids final $w_i = \sum_{k \in U_i^B} w'_{ik} \Big/ M_i^{AB}$.
- **Étape 4** Enfin, nous posons $w_{ik} = w_i$ pour tous les $k \in U_i^B$.

On pourrait envisager de calculer la probabilité de sélection π_{ik}^B de l'individu k du ménage i de S^B. Cette probabilité correspond à la probabilité de sélectionner n'importe lequel des M_i^B individus du ménage i, et il faut donc ainsi connaître chacune des M_i^B probabilités π_k^A du ménage i de S^B. Malheureusement, notamment dans le cas des enquêtes à plusieurs degrés, les probabilités π_k^A sont souvent inconnues. De plus, à part des cas relativement simples (par exemple, lorsque les individus k sont sélectionnés dans S^A de façon indépendante), le calcul des poids π_{ik}^A peut s'avérer très complexe. La *méthode du partage des poids* offre donc une solution simple à un problème de pondération difficile, voire impossible, à effectuer en pratique.

4.2 Une généralisation du problème

Imaginons des liens (ou correspondance) unissant les individus des deux vagues d'enquête. Puisqu'il s'agit de suivre des individus au cours du temps, ces liens peuvent être vus comme étant de « un à un » (figure 4.1). Lors de discussions avec Jean-Claude Deville, ce dernier a eu l'idée suivante : « Pourquoi ne pas généraliser les liens ? », c'est-à-dire, au lieu de liens « un à un », pourquoi pas considérer des liens « plusieurs à plusieurs » (figure 4.2) ? Les figures 4.1 et 4.2 contiennent une représentation graphique des méthodes. L'échantillon S^A est le sous-ensemble en jaune de la vague A. Les sous-ensembles en vert de la vague B sont les grappes U_i^B (les ménages) existant lors de la deuxième vague.

Figure 4.1 Liens longitudinaux (« un à un »).

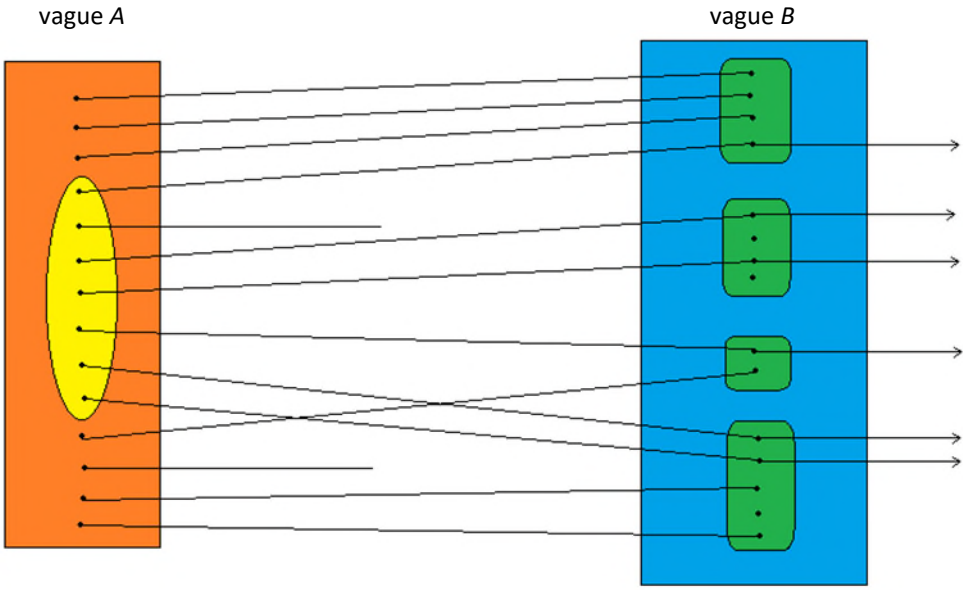
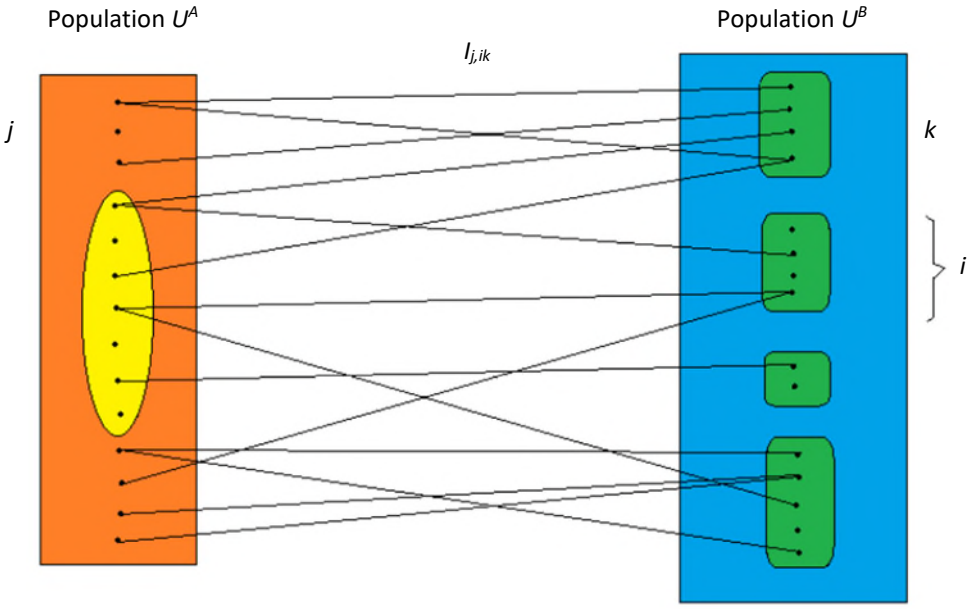


Figure 4.2 Liens quelconques (« plusieurs à plusieurs »).



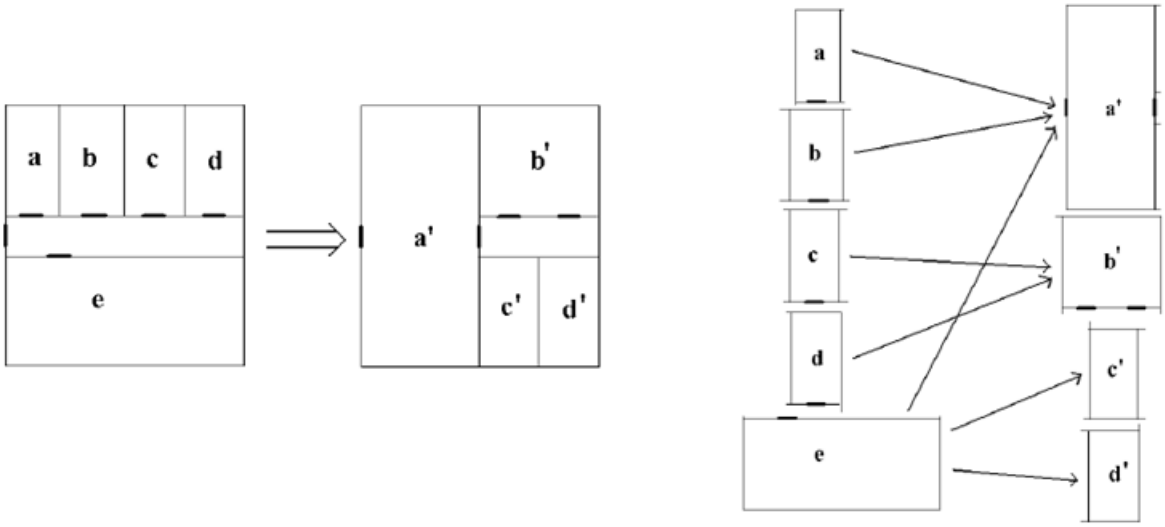
Avec cette nouvelle façon de voir les liens, la question devenait alors : « comment associer un poids (sans biais) aux unités enquêtées de U^B (la population cible) à la suite de la sélection d'unités dans U^A (la base de sondage) ? ». En fait, le problème s'élargissait à un contexte beaucoup plus large que celui des enquêtes longitudinales.

Le nouveau problème étudié était alors le suivant. Soit deux populations U^A et U^B reliées entre elles. On désire produire une estimation pour U^B (population cible), mais une base de sondage n'est disponible que pour U^A. La solution proposée est alors de tirer un échantillon de U^A afin de produire des estimations pour U^B en se servant de la correspondance (liens) existante entre les deux populations. Pour cette nouvelle façon de procéder, Jean-Claude Deville a alors proposé le terme sondage indirect.

Le processus d'enquête du sondage indirect est le suivant. Tout d'abord, pour chaque unité j de S^A, on identifie les unités k des grappes i de U^B qui ont un lien avec j. Soit U_i^B, l'ensemble des unités k de la grappe i. Pour chaque unité k identifiée, on établit la liste des M_i^B unités de la grappe i contenant cette unité. Finalement, on enquête auprès de toutes les unités k des grappes $i \in S^B$ pour mesurer la variable d'intérêt y.

Pour illustrer le sondage indirect, Jean-Claude Deville a proposé un exemple où on cherche à effectuer une enquête auprès de personnes (unités) qui habitent des logements (grappes). On dispose pour cela d'une base de sondage de logements, mais qui n'est malheureusement pas à jour. Cette base de sondage ne contient pas, entre autres, les rénovations touchant les divisions des logements des immeubles. Un exemple de ce type de rénovation est illustré à gauche de la figure 4.3. On remarque que les logements a, b, c, d et e ont été transformés pour obtenir les logements a', b', c' et d'. En tirant un échantillon de logements de la base de sondage, on se rend alors aux nouveaux logements en utilisant la correspondance entre les anciens et les nouveaux logements. Cette correspondance est illustrée à droite de la figure 4.3.

Figure 4.3 Logements avant et après rénovation (gauche) et sondage indirect des logements rénovés (droite).



L'estimation du total t_y^B de la population cible U^B peut se faire en se servant de S^A tiré de U^A. On note cependant que ceci peut constituer un défi de taille si les liens entre les unités de U^A et U^B ne sont pas bijectifs. En fait, dans ce cas, il s'avère difficile, voire impossible, d'associer une probabilité de sélection, ou un poids d'estimation, aux unités enquêtées dans U^B. La solution consiste alors à utiliser la méthode généralisée du partage des poids (MGPP) qui permet d'obtenir un poids d'estimation pour chaque unité enquêtée de la population cible U^B.

Tout comme dans le cas des enquêtes longitudinales, on suppose un échantillon S^A contenant m^A unités tirées de U^A contenant M^A unités. La population cible U^B contient M^B unités et celle-ci est divisée en N^B grappes, où la grappe i contient M_i^B unités. Les liens (ou correspondance) entre les unités j de U^A et les unités k des grappes i de U^B sont identifiés par la variable $l_{i,jk}$, où $l_{j,ik} = 1$ s'il existe un lien entre l'unité $j \in U^A$ et l'unité k de la grappe i de U^B, et 0 sinon.

Pour appliquer la MGPP, il suffit de suivre ces étapes (qui rappellent celles de la méthode du partage des poids, mais en plus générales) :

- **Étape 1** Pour chaque unité k des grappes i de S^B, on calcule le poids initial $w'_{ik} = \sum_{j \in U^A} l_{j,ik}\, \gamma_j \big/ \pi_j^A$, où $\gamma_j = 1$ si $j \in S^A$, et 0 sinon.
- **Étape 2** Pour chaque unité k des grappes i de S^B, on obtient le nombre total de liens $L_{ik}^B = \sum_{j \in U^A} l_{j,ik}$.
- **Étape 3** On calcule le poids final $w_i = \sum_{k \in U_i^B} w'_{ik} \Big/ \sum_{k \in U_i^B} L_{ik}^B$.
- **Étape 4** Enfin, nous posons $w_{ik} = w_i$ pour tous les $k \in U_i^B$.

Lavallée (2002, 2007) mentionne que le sondage indirect et la MGPP sont utiles parce qu'ils proposent une solution simple à des problèmes de sondage et de pondération complexes. De plus, en général, la MGPP donne les mêmes résultats que la théorie classique pour les problèmes simples. En fait, la MGPP constitue une solution intéressante, même si elle n'est pas toujours la plus précise (variance minimale) par rapport à une autre méthode d'estimation plus complexe.

4.3 Les propriétés de la MGPP

La détermination des propriétés de la MGPP s'est déroulée au cours de discussions avec Jean-Claude Deville qui ont débuté en 1995. Celles-ci ont mené au théorème et à ses deux corollaires suivants :

Théorème 1 *Dualité de la forme de $\hat{t}_y^B$ par rapport à U^A et U^B. L'estimateur $\hat{t}_y^B$ peut s'écrire sous les deux formes :*

$$\hat{t}_y^B = \sum_{i \in S^B} \sum_{k \in U_i^B} w_{ik} y_{ik}$$

(avec les poids de la MGPP) et

$$\hat{t}_y^B = \sum_{j \in U^A} \gamma_j Z_j \Big/ \pi_j^A ,$$

où $Z_j = \sum_{i \in U^B} \sum_{k \in U_i^B} l_{j,ik} Y_i \Big/ L_i^B$.

Le Théorème 1 nous montre que l'on est, finalement, en présence d'un simple estimateur de Horvitz-Thompson. De cette constatation découlent les deux corollaires suivants :

Corollaire 1 *Biais de* $\hat{t}_y^B$. *L'estimateur* $\hat{t}_y^B$ *est sans biais pour l'estimation de* Y^B, *par rapport au plan de sondage.*

Corollaire 2 *Variance de* $\hat{t}_y^B$. *La formule de la variance de l'estimateur* $\hat{t}_y^B$, *par rapport au plan de sondage, est donnée par*

$$V_p(\hat{t}_y^B) = \sum_{j \in U^A} \sum_{j' \in U^A} \left(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A \right) \frac{Z_j Z_{j'}}{\pi_j^A \pi_{j'}^A},$$

où $\pi_{jj'}^A$ *désigne la probabilité conjointe de sélection des unités* j *et* j'.

4.4 Le calage

Supposons que l'on désire corriger les poids de la MGPP pour que les estimations produites correspondent à des totaux connus (information auxiliaire). La technique la plus utilisée est alors le calage développé par Deville et Särndal (1992).

Dans le contexte du sondage indirect, il y a deux sources possibles d'information auxiliaire :

(i) de la base de sondage U^A, on peut disposer d'un vecteur colonne $\mathbf{x}_j^A$ et de son total $\mathbf{t}_x^A = \sum_{j \in U^A} \mathbf{x}_j^A$ (supposé connu);

(ii) de la population cible U^B, on peut aussi disposer d'un vecteur colonne $\mathbf{x}_{ik}^B$ et de son total $\mathbf{t}_x^B = \sum_{i \in U^B} \sum_{k \in U_i^B} \mathbf{x}_{ik}^B$ (supposé connu).

Les contraintes de calage associées à la MGPP sont :

(i) $\hat{\mathbf{t}}_x^{\mathrm{CAL},A} = \sum_{j \in S^A} w_j^{\mathrm{CAL},A} \mathbf{x}_j^A = \mathbf{t}_x^A$ et

(ii) $\hat{\mathbf{t}}_x^{\mathrm{CAL},B} = \sum_{i \in S^B} \sum_{k \in U_i^B} w_{ik}^{\mathrm{CAL},B} \mathbf{x}_{ik}^B = \mathbf{t}_x^B$, où $w_j^{\mathrm{CAL},A}$ est le poids de calage obtenu à partir des $d_j^A = 1 \big/ \pi_j^A$, et $w_{ik}^{\mathrm{CAL},B}$ est le poids de calage de l'unité k de la grappe enquêtée i où on a appliqué la MGPP.

À partir du Théorème 1, on peut réécrire cette dernière contrainte sous la forme : $\hat{\mathbf{t}}_x^{\mathrm{CAL},B} = \sum_{j \in S^A} w_j^{\mathrm{CAL},A} \mathbf{\Gamma}_j = t_x^B$, où $\mathbf{\Gamma}_j = \sum_{i \in U^B} \sum_{k \in U_i^B} l_{j,ik} X_i^B \Big/ L_i^B$. On voit alors que cette contrainte est maintenant exprimée en fonction des unités $j \in s^A$.

En définissant les vecteurs

$$\mathbf{x}_j^{AB} = \begin{pmatrix} \mathbf{x}_j^A \\ \boldsymbol{\Gamma}_j \end{pmatrix} \quad \text{et} \quad \mathbf{t}_x^{AB} = \begin{pmatrix} \mathbf{t}_x^A \\ \mathbf{t}_x^B \end{pmatrix},$$

on obtient alors une contrainte unique englobant U^A et U^B :

$$\hat{\mathbf{t}}_x^{\mathrm{CAL},AB} = \sum_{j \in S^A} w_j^{\mathrm{CAL},A} \mathbf{x}_j^{AB} = \mathbf{t}_x^{AB}.$$

La formulation de la détermination de l'estimateur $\hat{t}_y^{\mathrm{CAL},B} = \sum_{j \in S^A} w_j^{\mathrm{CAL},A} Z_j$ associé à la MGPP est : Déterminer $w_j^{\mathrm{CAL},A}$, pour $j \in S^A$, afin de minimiser la distance totale

$$\sum_{j \in S^A} G_j(w_j^{\mathrm{CAL},A}, d_j^A)$$

sous la contrainte unique

$$\hat{\mathbf{t}}_x^{\mathrm{CAL},AB} = \sum_{j \in S^A} w_j^{\mathrm{CAL},A} \mathbf{x}_j^{AB} = \mathbf{t}_x^{AB}.$$

Cette formulation correspond à la formulation du calage de Deville et Särndal (1992). Le calage peut donc s'appliquer relativement aisément au sondage indirect et la MGPP.

Il est important de noter que ce travail sur le calage a été réalisé sans collaboration directe avec Jean-Claude Deville. Cependant, après tout ce travail présenté par Lavallée (2001) au Colloque francophone sur les sondages à Bruxelles, ce dernier a découvert l'article de Deville (1998b) où on retrouve la même solution au calage associé à la MGPP.

4.5 Optimisation des liens

La variable indicatrice $l_{j,ik}$ indique s'il y a un lien ou non entre les unités j de la base de sondage U^A et les unités k des grappes i de la population cible U^B. Cependant, elle n'indique pas l'importance relative que pourraient avoir certains liens par rapport à d'autres. Il est possible de remplacer $l_{j,ik}$ par une variable quantitative $\theta_{j,ik}$ représentant l'importance qu'on veut donner au lien $l_{j,ik}$. Cette variable $\theta_{j,ik}$ est définie sur $[0,+\infty)$, où $\theta_{j,ik} = 0$ équivaut à $l_{j,ik} = 0$. Il est à noter que si le processus d'assignation des valeurs de $\theta_{j,ik}$ est indépendant du tirage de S^A, la MGPP reste sans biais.

En remplaçant les liens $l_{j,ik}$ par $\theta_{j,ik}$, on obtient un nouvel estimateur (sans biais) $\hat{t}_{y\theta}^B$. Le problème qui nous intéresse alors est de déterminer des valeurs optimales de $\theta_{j,ik}$ de manière à minimiser la variance de $\hat{t}_{y\theta}^B$. En effet, puisque l'estimateur $\hat{t}_{y\theta}^B$ reste sans biais quelles que soient les valeurs de $\theta_{j,ik}$, on doit pouvoir déterminer des valeurs de ces dernières pour maximiser la précision de $\hat{t}_{y\theta}^B$. Le problème est donc de déterminer $\left[\theta_{j,ik}\right]_{M^A \times M^B}$ afin de minimiser $V_p(\hat{t}_{y\theta}^B) = f(y_{ik};\ i = 1, \ldots, N^B;\ k = 1, \ldots, M_i^B)$.

Deville et Lavallée (2006) ont déterminé les valeurs de $\theta_{j,ik}$ telles que la variance de l'estimateur $\hat{t}_{y\theta}^B$ soit (presque) minimale. La solution optimale n'est pas simple à écrire, et elle dépend souvent de la variable d'intérêt y. Jean-Claude Deville a cependant eu l'idée de définir une optimalité faible ainsi qu'une optimalité forte indépendante des y.

L'optimalité faible consiste à déterminer des valeurs de $\theta_{j,ik}$ afin de minimiser la variance de $\hat{t}_{y\theta}^{B}$ pour un choix très précis d'une variable d'intérêt : $y_{ik}=1$ pour une unité k donnée d'une grappe i de U^B et $y_{i'k'}=0$ pour toutes les autres unités de U^B. Le problème d'optimisation se réduit alors à déterminer $[\theta_{j,ik}]_{M^A\times M^B}$ afin de minimiser $V_p(\hat{t}_{y\theta}^{B})=f(y_{ik}=1;\ y_{i'k'}=0;\ \forall i\neq i'$ et $k\neq k')$. Deville et Lavallée (2006) précisent que l'optimalité faible correspond à minimiser la variance $\mathrm{V}(w_{ik}^{\theta}\,|\,ik\in S^B)$ du poids w_{ik}^{θ} obtenu par la MGPP (avec $\theta_{j,ik}$ au lieu de $l_{j,ik}$) pour toutes les valeurs possibles de $[\theta_{j,ik}]_{M^A\times M^B}$. On note que les liens pondérés faiblement optimaux résultants ne font pas intervenir la variable y proprement dite (puisque les valeurs de y ont été remplacées par 1 et 0). De plus, les valeurs faiblement optimales des $\theta_{j,ik}$ sont généralement relativement faciles à calculer.

L'optimalité forte indépendante de y fait intervenir une étape supplémentaire à l'optimalité faible. En effet, il s'agit de vérifier que les valeurs faiblement optimales des $\theta_{j,ik}$ ne dépendent pas, de manière générale (c'est-à-dire pour toutes valeurs de y autres que 1 et 0) de la variable d'intérêt y. Dans cet esprit, Deville et Lavallée (2006) ont proposé un critère pour vérifier si l'optimalité faible correspond à l'optimalité forte (variance minimale de $\hat{t}_{y\theta}^{B}$). Si ce critère est satisfait, l'optimalité forte ne dépend alors pas de la variable d'intérêt y.

5. Le développement de l'expression de la variance et son estimation pour des estimateurs complexes

Le développement de l'expression de la variance et son estimation pour un estimateur obtenu par sondage constitue une étape essentielle pour produire les intervalles de confiance qui renseigneront les utilisateurs des statistiques sur leur degré de fiabilité. La théorie classique utilise soit une approche analytique, par nature à base de formules mathématiques, soit une approche par réplication d'échantillons (bootstrap, jackknife, random groups). Grossièrement, on peut considérer que l'approche analytique s'applique plutôt lorsque l'échantillonnage est compliqué et l'estimateur d'expression plutôt simple, alors qu'on utilise l'approche par réplication plus volontiers dans la configuration inverse, c'est-à-dire en présence d'un échantillonnage simple et d'un estimateur complexe. C'était en tout cas une stratégie couramment appliquée avant que l'on ne développe la théorie de la linéarisation des statistiques complexes. Ce développement doit beaucoup à Jean-Claude Deville. Bien sûr, on connaissait depuis de nombreuses années la méthode de linéarisation des estimateurs définis comme des fonctions de composantes linéaires, typiquement des fonctions dérivables d'estimateurs de totaux comme un ratio ou un coefficient de régression ou d'estimateurs qui sont la solution d'équations estimantes (Woodruff, 1971; Binder, 1983; Wolter, 1985; Binder, 1991; Francisco et Fuller, 1991; Binder et Kovačević, 1995; Binder, 1996). À la fin des années 90, Jean-Claude Deville a proposé un cadre formel basé sur la fonction d'influence dans la revue *Techniques d'enquêtes* (Deville, 1999) pour traiter dans un cadre asymptotique mais en toute généralité les statistiques hautement non-linéaires, dont par exemple les fractiles ou des paramètres définis comme solutions de certaines équations (paramètres implicites). Lorsque la taille d'échantillon est grande, la linéarisation permet

finalement d'approcher un estimateur très complexe par un estimateur linéaire classique de type Horvitz-Thompson, puis la variance du premier par celle du deuxième, produisant ainsi le résultat recherché.

Plus précisément, l'approche historique considère le paramètre θ et l'estimateur $\hat{\theta}$ comme des fonctions dérivables des variables d'intérêt individuelles. La linéarisation s'appuie alors sur un développement en série de Taylor de $\hat{\theta}$ autour de son espérance θ. Les poids de sondage sont présents dans l'expression $\hat{\theta}$ mais ils ne sont pas considérés comme des variables. Jean-Claude Deville renverse d'une certaine façon l'approche en considérant $\hat{\theta}$ comme une fonction des poids de sondage et mobilise une forme de dérivée par rapport à ces poids : c'est la fonction d'influence, introduite dans la partie suivante.

5.1 Le cadre théorique

La méthodologie proposée s'expose en trois étapes : *primo* un cadre asymptotique, *secundo* une formalisation utilisant la notion de mesure sur un espace probabiliste, et *tertio* l'exploitation du concept de fonction d'influence, opportunément adapté au contexte.

Le cadre asymptotique est celui défini dans Isaki et Fuller (1982), et considère une série de populations emboitées, de tailles respectives N tendant vers l'infini et au sein desquelles on tire des échantillons s dont la taille n tend également vers l'infini. Pour toute variable individuelle x_k, si t_x est le vrai total des x_k et

$$\hat{t}_x = \sum_{k \in s} w_k x_k$$

son estimateur, alors on convient que $N^{-1}t_x$ a une limite, que $N^{-1}(\hat{t}_x - t_x)$ tend vers 0 en probabilité et que $\sqrt{n}N^{-1}(\hat{t}_x - t_x)$ converge en loi vers une loi de Gauss. Toute statistique S plus complexe construite à partir de totaux – vrais ou estimés – relève d'hypothèses de même nature mais étendues : selon son expression, dans les cas les plus fréquents la convergence survient lorsqu'elle est multipliée par $N^{-\alpha}$ où α est un entier positif. On parle d'homogénéité de degré α : un ratio est une statistique homogène de degré 0, une variance une statistique homogène de degré 2. Ainsi, le premier axiome s'étend en convenant que $N^{-\alpha}S$ a une limite.

Intervient ensuite la formalisation des estimateurs en mobilisant la notion de mesure. Dans l'expression de tout paramètre, les individus de la population finie sont « naturellement » pondérés par un poids égal à 1, interprété comme une masse associée à une mesure M de support fini. Dans cette même population, un échantillonnage conduit à pondérer tout individu k de l'échantillon tiré s par un poids de sondage w_k et tout individu k hors de s par 0. Ce qui définit une mesure $\hat{M}$. Un paramètre, aussi complexe soit-il, peut s'exprimer comme une fonction de M, notée $T(M)$ et appelée « fonctionnelle de M ». Considérant un total par exemple,

$$T(M) = \sum_{k=1}^{N} y_k = \sum_{k=1}^{N} y_k \, M(k).$$

Adoptant une notation générale familière en théorie de la mesure, on écrira $T(M) = \int y dM$, l'intégration portant sur l'ensemble des individus de la population. Au niveau des estimateurs, toujours dans le cas d'un total, on considérera

$$T(\hat{M}) = \sum_{k \in s} w_k y_k = \sum_{k=1}^{N} y_k \hat{M}(k),$$

auquel cas $T(\hat{M}) = \int y d\hat{M}$. Ce parallélisme s'applique pour tout paramètre complexe, exprimé initialement sous forme $T(M)$ et estimé par $T(\hat{M})$, estimateur obtenu par substitution de $\hat{M}$ à M sur la population finie.

La troisième composante de la théorie utilise la notion de fonction d'influence, utilisée sous une forme voisine en théorie de la robustesse (Hampel, Ronchetti, Rousseeuw et Stahel, 1985). On considère la mesure spécifique δ_k obtenue en affectant la masse 1 à l'individu k puis la mesure $M + t\delta_k$ affectant par construction la masse $1+t$ à l'individu k et la masse 1 à tous les autres individus. La fonction d'influence est définie – lorsque la limite existe – par

$$IT(M,k) = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}.$$

On peut montrer, sous certaines conditions techniques le plus souvent vérifiées dans les contextes rencontrés, que lorsque l'espace des mesures est doté d'une distance, si une mesure M_2 converge vers une mesure M_1 alors

$$T(M_2) = T(M_1) + \int IT(M_1, k)\, dM_2 - \int IT(M_1, k)\, dM_1 + R_\epsilon,$$

où R_ϵ est un résidu aléatoire qui tend en probabilité vers zéro. En la circonstance, il faut adapter cette égalité aux conditions asymptotiques initiales postulées : pour traiter de façon générale les statistiques homogènes de degré α, ce sont les fonctionnelles $N^{-\alpha}T$ qui ont les propriétés asymptotiques requises et qu'il faut donc ici considérer. Par ailleurs, remarquant que la masse totale associée à la mesure M est N, en définissant une distance entre métriques selon

$$d\left(\frac{M_1}{N}, \frac{M_2}{N}\right) = \left| \int y d\left(\frac{M_1}{N}\right) - \int y d\left(\frac{M_2}{N}\right) \right|$$

et en fixant $M_1 = M$ et $M_2 = \hat{M}$, les postulats asymptotiques traduisent la convergence de $\hat{M}/N$ vers M/N et conduisent finalement à :

$$N^{-\alpha} T(\hat{M}) = N^{-\alpha} T(M) + \int IT(M,k)\, d\left(\frac{\hat{M}}{N}\right) - \int IT(M,k)\, d\left(\frac{M}{N}\right) + R_\epsilon,$$

où le résidu R_ϵ est dans ces conditions négligeable en probabilité devant $1/\sqrt{n}$, c'est-à-dire que

$$\text{pour tout } \epsilon > 0, P\left(\left| \sqrt{n} R_\epsilon > \epsilon \right| \right) \to 0.$$

En notant $IT(M,k) = z_k$, il en découle

$$N^{-\alpha}(T(\hat{M}) - T(M)) = \frac{1}{N} \sum_{k \in s} w_k z_k - \frac{1}{N} \sum_{k=1}^{N} z_k + R_\epsilon. \tag{5.1}$$

En notant $\hat{t}_z = \sum_{k \in s} w_k z_k$ l'estimateur linéaire naturel du total $t_z = \sum_{k=1}^{N} z_k$,

$$\sqrt{n} N^{-\alpha} (T(\hat{M}) - T(M)) = \sqrt{n} \frac{1}{N} (\hat{t}_z - t_z) + R_\epsilon.$$

D'après le troisième postulat asymptotique, le terme de droite a une limite Gaussienne, et donc le terme de gauche a une variance limite, qui est égale à celle du membre de droite. Le pragmatisme habituel conduira alors à utiliser $N^{2(\alpha-1)} V(\hat{t}_z)$ comme variance approchée de $T(\hat{M})$ lorsqu'on considérera que n est « suffisamment grand ». La variable z_k est dite variable linéarisée associée à la fonctionnelle T. Pour procéder concrètement à l'estimation de la variance, à partir du moment où z_k dépend d'un nombre fini de paramètres que l'on est en mesure d'estimer en mobilisant les données de l'échantillon, on remplacera z_k par son estimateur naturel $\hat{z}_k$. La variance estimée diffère alors de la vraie variance par un terme dont l'ordre de grandeur est en $n^{-1/2}$.

L'article de Deville a encore une fois été novateur car il été suivi de plusieurs autres travaux sur l'estimation de variance. Alors que Deville a proposé de linéariser le paramètre d'intérêt au niveau de la population, Demnati et Rao (2004, 2010) dérivent directement l'estimateur par rapport aux poids de sondage. Cette méthode est une manière simple de calculer la fonction d'influence sur l'estimateur. Graf (2011), Antal, Langel et Tillé (2011), Graf et Tillé (2014) et Vallée et Tillé (2019) dérivent l'estimateur par les indicatrices d'appartenance à l'échantillon, ce qui permet de prendre en compte à la fois la non-linéarité de l'estimateur et le calage. Les résultats données par les différentes méthodes ne sont pas toujours identiques car les poids peuvent dépendre des indicatrices d'appartenance notamment quand l'estimateur est calé.

5.2 Les outils

De la théorie précédente, découlent des règles de calcul de variables linéarisées qui permettent de traiter plus simplement les estimateurs complexes, en les décomposant. La fonctionnelle « total » $T(M) = \sum_{i=1}^{N} y_i$ est la plus simple. Puisque

$$T(M + t\delta_k) = \sum_{i=1}^{N} y_i M(i) + \sum_{i=1}^{N} y_i t \delta_k(i) = \sum_{i=1}^{N} y_i + t y_k,$$

il est immédiat que $z_k = IT(M,k) = y_k$. L'expression (5.1) est ici tautologique, avec $R_\epsilon = 0$.

Différentes propriétés utiles sont citées dans l'article fondateur.

(i) Soit $T(M)$ et $S(M)$ deux fonctionnelles :

$$I(T+S)(M,k) = IT(M,k) + IS(M,k)$$

et

$$I(T \cdot S)(M,k) = IT(M,k) \cdot S(M,k) + IS(M,k) \cdot T(M,k).$$

(ii) Soit $T(M)$ un vecteur de totaux dans $\mathbb{R}^p$ et f une fonction différentiable de $\mathbb{R}^p$ dans $\mathbb{R}$ dont la matrice (p,p) des dérivées partielles prises en $T(M)$ est notée $D_{f,T(M)}$. Alors, $If(T)(M,k) = D_{f,T(M)} \cdot IT(M,k)$. Cette règle est utile pour retrouver les linéarisées, bien connues, des fonctions régulières de totaux, comme par exemple les ratios ou les coefficients de corrélation linéaire.

On s'intéresse maintenant à des fonctionnelles paramétrées par un vecteur α de $\mathbb{R}^p$, notées $T(M,\alpha)$. Par exemple, si $p=1$ et $\alpha \in [0,1]$, notant F la fonction de répartition associée à la distribution des y_k, $T(M,\alpha) = F^{-1}(\alpha)$ est le quantile d'ordre α de la distribution.

(iii) Si $T(M,\lambda)$ en tant que fonction de λ a une régularité suffisante, on peut définir une fonctionnelle réciproque $\Lambda(M,\alpha)$ vérifiant par conséquent $T(M,\Lambda(M,\alpha)) = \alpha$. Par exemple, pour traiter la variance d'un quantile on considérera $T(M,\lambda) = F(\alpha)$ et on obtiendra $\Lambda(M,\alpha) = F^{-1}(\alpha)$. On peut montrer que

$$I\Lambda(M,\alpha,k) = -\left\{\frac{\partial T}{\partial \alpha}(M,\Lambda(M,\alpha))\right\}^{-1} \cdot IT(M,\alpha,k). \tag{5.2}$$

Ce théorème est utile pour obtenir, par exemple, la linéarisée d'un quantile ou d'un paramètre implicite (voir 5.3).

(iv) Supposons que le paramètre s'écrive comme fonction de la valeur y, c'est-à-dire considérons une fonctionnelle du type $T(M,\phi(y))$ où ϕ est une fonction ayant les bonnes propriétés techniques, puis la fonctionnelle $S(M) = \int T(M,\phi(y))\,dM$. Alors,

$$IS(M,k) = T(M,\phi(y_k)) + \int IT(M,\phi(y),k)\,dM. \tag{5.3}$$

Ce théorème sert, par exemple, pour déterminer la linéarisée du coefficient de Gini (voir 5.3).

(v) Si le paramètre est lui-même une fonctionnelle $S(M)$, on peut obtenir la fonction d'influence de $T(M,S(M))$ – c'est-à-dire la linéarisée d'une composée de fonctionnelles – selon

$$IT(M,S(M),k) = IT(M,\alpha,k) + \frac{\partial T}{\partial \alpha}(M,\alpha) \cdot IS(M,k),$$

où α prend dans l'expression finale de droite la valeur $S(M)$.

On cite enfin une propriété intéressante : pour tout fonctionnelle de degré α, on a

$$\sum_{k=1}^{N} IT(M,k) = \alpha \cdot T(M).$$

En particulier si $\alpha = 0$ (un ratio par exemple), la somme sur la population des variables linéarisées z_k est nulle.

5.3 Quelques applications

La théorie qui précède permet de linéariser (à peu près) tous les estimateurs que l'on est en mesure de rencontrer dans les opérations statistiques d'enquête. Elle a donc une portée extrêmement générale, et permet *in fine* tous les calculs analytiques de variance, pour (à peu près) tous les paramètres imaginables, dès lors qu'on accepte les conditions asymptotiques, c'est-à-dire que l'on considère que la taille d'échantillon n est « suffisamment grande ».

L'article originel de Jean-Claude Deville présente plusieurs exemples d'applications pour des paramètres complexes. On y trouve, avec les développements techniques qui les justifient, les cas des paramètres implicites, des quantiles, du coefficient de Gini, du seuil de pauvreté (défini comme la proportion d'individus dans une population dont la variable revenu est inférieure à la moitié de sa médiane), le coefficient de corrélation des rangs de Kendall. Sont également traitées l'estimation de variance d'une composante principale d'un nuage de points, et celle de la projection d'un point représentant une sous-population quelconque sur un axe factoriel dans le cadre d'une analyse des correspondances multiples. Ce qui suit fournit quelques résultats concernant les paramètres implicites, les quantiles, et enfin un coefficient d'inégalité de type Gini.

Un paramètre implicite est un vecteur de $\mathbb{R}^p$ solution d'une équation du type $H(M,\mathbf{B})=0$ où $H(M,\mathbf{B})=\sum_{k\in U} l_k(\mathbf{B})$, les fonctions l_k étant des fonctions régulières de $\mathbb{R}^p$ dans $\mathbb{R}^p$. En utilisant (5.2), et en notant $\mathbf{B}_0$ la solution de l'équation, on obtient pour tout k de U,

$$I\mathbf{B}(M,k) = -\left\{\frac{\partial H}{\partial \mathbf{B}}(M,\mathbf{B}_0)\right\}^{-1} \cdot IH(M,\mathbf{B}_0,k)$$

c'est-à-dire

$$I\mathbf{B}(M,k) = -\left\{\sum_{k\in U}\frac{\partial l_k}{\partial \mathbf{B}}(M,\mathbf{B}_0)\right\}^{-1} \cdot l_k(\mathbf{B}_0).$$

Cette situation est par exemple celle des coefficients de régression, les l_k découlant des équations normales. Si la régression est linéaire, on en tire (notations classiques)

$$l_k(\mathbf{B}) = \frac{1}{\sigma_k^2}\cdot \mathbf{x}_k(y_k - \mathbf{z}_k^\top \mathbf{B})$$

et finalement le vecteur linéarisé de $\mathbb{R}^p$

$$\mathbf{z}_k = I\mathbf{B}(M,k) = -\left(\sum_{k\in U}\frac{\mathbf{x}_k\mathbf{z}_k^\top}{\sigma_k^2}\right)^{-1}\cdot\frac{1}{\sigma_k^2}\cdot\mathbf{x}_k(y_k-\mathbf{z}_k^\top\mathbf{B}_0)$$

que l'on peut trouver aussi par l'approche « historique ». Si la régression est logistique, les outils classiques ne suffisent plus et alors

$$\mathbf{z}_k = I\mathbf{B}(M,k) = \left\{\sum_{k\in U}\mathbf{x}_k\mathbf{z}_k^\top \cdot f(\mathbf{z}_k^\top\mathbf{B}_0)(1-f(\mathbf{z}_k^\top\mathbf{B}_0))\right\}^{-1}\mathbf{x}_k(y_k - f(\mathbf{z}_k^\top\mathbf{B}_0)),$$

où $f(u) = e^u(1+e^u)$.

Pour obtenir la linéarisée d'un quantile, il faut considérer la fonctionnelle « fonction de répartition »

$$F(x) = T(M,x) = \frac{1}{N}\cdot \text{Card}\{k\in U/x_k \leqslant x\} = \frac{1}{N}\cdot\int 1_{x_k\leqslant x}\,dM,$$

dont la fonction d'influence vaut

$$IT(M,x,k) = \frac{1}{N}\cdot\{1_{x_k\leqslant x} - F(x)\}.$$

En supposant pour simplifier que cette fonction croissante est dérivable et inversible, on définirait le quantile q_α d'ordre α où $\alpha\in[0,1]$ par l'équation $F(q_\alpha)=\alpha$. L'application de la formule (5.2) conduirait alors à la linéarisée

$$z_k = Iq_\alpha(M,k) = -[N\cdot F'(q_\alpha)]^{-1}\cdot(1_{x_k\leqslant q_\alpha} - \alpha).$$

Jean-Claude Deville a proposé un aménagement astucieux et assez simple pour tenir compte de la forme « en escalier » de $F(x)$, qui n'est en réalité ni dérivable ni inversible.

Terminons en précisant la linéarisée d'un indice d'inégalité type « indice de Gini ». L'indice considéré dans l'article est défini par

$$\text{GINI} = \frac{1}{t_x}\int xF(x)\,dM,$$

où $t_x = \sum_{k\in U} x_k$. En utilisant (5.3) on obtient

$$z_k = F(x_k)\cdot\frac{x_k - \overline{x}_{k,\text{inf}}}{t_x} - \text{GINI}\cdot\frac{x_k}{t_x},$$

en posant

$$\overline{x}_{k,\text{inf}} = \frac{\int x1_{x<x_k}\,dM}{\int 1_{x<x_k}\,dM}$$

qui représente la moyenne des x inférieurs à x_k.

6. L'échantillonnage par quotas

L'échantillonnage « par quotas » est une méthode de sondage peu pratiquée par les instituts nationaux de statistique. L'argument essentiel est celui du biais des estimateurs qui en découlent, alors même que l'on

ne peut pas se prévaloir de variances plus faibles qu'avec une méthode probabiliste bien choisie, compte tenu des grandes tailles d'échantillon communément gérées par la statistique publique. Un autre argument, plus philosophique, est la dépendance de la qualité de l'estimation envers un modèle, c'est-à-dire un jeu d'hypothèses simplificatrices de la réalité, parfois même (beaucoup) trop simplificatrices. Jean-Claude Deville avançait souvent ce second argument, qu'il considérait comme traduisant une forme d'absence de neutralité qui ne convenait pas à un institut national de statistique, du moins quand il est possible de faire autrement (on peut rétorquer que l'on utilise systématiquement des modèles pour traiter la non-réponse, mais c'est incontournable!). C'est peut-être parce que cette méthode de sondage empirique très utilisée (particulièrement dans le secteur privé) est par nature risquée et donc controversée que Jean-Claude Deville a éprouvé le besoin de formaliser la question. Il semble qu'il ait été le premier à le faire de manière aussi complète, rédigeant un article faisant autorité dans la revue *Techniques d'enquêtes* en 1991 (Deville, 1991). Dans cet article, deux types de modèles sont distingués : un modèle porte sur l'échantillonnage, un autre sur la variable d'intérêt. Dans chaque configuration, un estimateur est explicité, son biais étudié, et lorsque c'est possible l'auteur produit une expression théorique de variance ainsi qu'un estimateur sans biais de cette variance.

6.1 Le cadre général

Rappelons le principe du sondage par quotas, en se restreignant pour simplifier au cas d'une ou de deux variables dites « de quota ». L'information auxiliaire est constituée par les tailles des sous-populations définies par les modalités des variables de quota, qui sont donc de nature qualitative. S'il y a une seule variable qualitative à I modalités pour définir ces sous-populations, on dispose de l'effectif N_i pour chaque modalité i variant de 1 à I. S'il y a deux variables qualitatives possédant respectivement I et J modalités, notant $N_{i,j}$ l'effectif de la cellule (i,j), on connaît les effectifs marginaux $N_{i,.} = \sum_{j=1}^{J} N_{i,j}$ pour tout i variant de 1 à I ainsi que les effectifs marginaux $N_{.,j} = \sum_{i=1}^{I} N_{i,j}$ pour tout j variant de 1 à J (quotas « marginaux »). Noter que la connaissance des effectifs croisés $N_{i,j}$ (quotas « croisés ») nous ramène au cas d'une seule variable qualitative. La sélection de l'échantillon se fait de manière empirique, sans base de sondage, en respectant quelques consignes de collecte pour aléatoiriser autant que possible la composition de l'échantillon de taille globale n tout en imposant des contraintes fixant les n_i (cas d'une variable de quota) ou affectant les $n_{i,j}$ (cas de deux variables de quota). Ces contraintes, que l'on nommera « contraintes de quota », sont fixées à l'appréciation du sondeur, mais en pratique on trouve presque exclusivement le contexte des quotas proportionnels, où il s'agit de faire en sorte que l'échantillon ait la même structure que celle de la population, soit $n_i = n\frac{N_i}{N}$ dans le cas d'une seule variable ou $n_{i,.} = n\frac{N_{i,.}}{N}$ pour tout i et $n_{.,j} = n\frac{N_{.,j}}{N}$ pour tout j dans le contexte des quotas croisés (contraintes « de quotas proportionnels »). Cette situation constitue un standard rassurant mais ne correspond pas à la situation intuitive d'optimum, car il est toujours préférable d'augmenter les tailles d'échantillon dans les cellules où la dispersion est la plus forte (se référer à l'optimum de Neyman dans le cas probabiliste stratifié – qui diffère de l'allocation proportionnelle tout en lui étant préférable).

6.2 Un modèle d'échantillonnage

Le cas des quotas sur une seule variable peut se traiter en assimilant le tirage à du tirage stratifié avec sondage aléatoire simple dans chaque strate, chaque sous-population associée à une modalité i de la variable de quota définissant une strate au sein de laquelle la taille d'échantillon n_i est imposée. On imagine difficilement d'autres alternatives et il n'y a donc rien de bien original à ajouter dans ce contexte simple.

Le cas intéressant est celui des quotas marginaux. Le modèle comporte deux temps. Dans un premier temps, on assimile le tirage à un tirage aléatoire simple sous contrainte, les contraintes étant celles imposées sur les $n_{i,\cdot}$ et sur les $n_{\cdot,j}$, donc les contraintes de quotas. C'est une hypothèse assez audacieuse, postulant une neutralité parfaite des enquêteurs lors de la sélection de leurs échantillons respectifs. Techniquement, le tirage aléatoire simple sous contrainte revient à considérer que tout échantillon ne respectant pas les quotas a une probabilité de sélection nulle et que tous les échantillons respectant les quotas ont la même probabilité de sélection. Pratiquement, on pourrait la mettre en œuvre en effectuant des sondages aléatoires simples successifs et indépendants, tout en les refusant tant que les contraintes de quota ne sont pas vérifiées (tirage réjectif). Mais dans un second temps, on oublie en quelque sorte l'existence de ces contraintes…

Si on note $P_{i,j}$ le poids $\frac{N_{i,j}}{N}$ de la cellule (i,j), l'objectif consiste à estimer ces poids. En effet, la moyenne $\overline{Y}$ de toute variable d'intérêt Y définie dans la population s'écrit $\overline{Y} = \sum_{(i,j)} \frac{N_{i,j}}{N} \cdot \overline{Y}_{i,j}$ et sous le modèle convenu la vraie moyenne $\overline{Y}_{i,j}$ dans la cellule (i,j) est estimée sans biais par la moyenne simple $\overline{y}_{i,j}$ dans l'échantillon recoupant cette cellule, si bien que l'estimateur naturel par quotas de $\overline{Y}$ sera $\hat{\overline{Y}}_{\text{quota}} = \sum_{(i,j)} \hat{P}_{i,j} \cdot \overline{y}_{i,j}$ où $\hat{P}_{i,j}$ estime $P_{i,j}$ – autant que possible sans biais. Avec un tirage aléatoire simple standard, les tailles d'échantillon par cellule (i,j), soit $n_{i,j}$, suivent une loi multinomiale. Le second temps du modèle intervient ici pour postuler que sous les contraintes de quota, la loi des $n_{i,j}$ reste mutinomiale. Dans ce contexte, l'estimation des $P_{i,j}$ se fait par exemple par maximum de vraisemblance, ce qui conduit à maximiser l'objectif $\prod_{(i,j)} P_{i,j}^{n_{i,j}}$. Les solutions doivent être compatibles avec l'information marginale connue, ce qui impose par ailleurs

$$\sum_{j=1}^{J} P_{i,j} = \frac{N_{i,\cdot}}{N} \quad \text{pour tout} \quad i \quad \text{et} \quad \sum_{i=1}^{I} P_{i,j} = \frac{N_{\cdot,j}}{N}, \quad \text{pour tout} \quad j.$$

Les solutions obtenues sont de la forme $\hat{P}_{i,j} = \frac{n_{i,j}}{n}(a_i + b_j)^{-1}$ (les coefficients inconnus a_i et b_j sont les coefficients de Lagrange associées à l'optimisation sous contrainte). Comme il y a $I + J - 1$ contraintes indépendantes, en imposant par exemple la contrainte identifiante $b_J = 0$, on obtient les a_i et les b_j en résolvant le système (S)

$$\sum_{j=1}^{J} n_{i,j}(a_i + b_j)^{-1} = n\frac{N_{i,\cdot}}{N}, \quad \text{pour tout} \quad i = 1,\ldots,I,$$
$$\sum_{i=1}^{I} n_{i,j}(a_i + b_j)^{-1} = n\frac{N_{\cdot,j}}{N}, \quad \text{pour tout} \quad j = 1,\ldots,J-1.$$

On constate à ce stade que si les quotas sont proportionnels (on rappelle qu'il s'agit du scénario à peu près systématiquement choisi en pratique), manifestement on résout le système en choisissant

systématiquement $a_i = 1$ et $b_j = 0,$ auquel cas $\hat{P}_{i,j} = \frac{n_{i,j}}{n}$ et $\hat{\bar{Y}}_{\text{quota}} = \bar{y}$ moyenne simple dans l'échantillon. On retrouve un résultat bien connu. Le cas des quotas non proportionnels fournit pour sa part un estimateur complexe, dont la forme analytique n'est pas explicite, mais asymptotiquement non biaisé dès lors que la pertinence du modèle – non discutée dans l'article fondateur – n'est pas remise en cause. Ce point doit inciter à la prudence parce que si les contraintes de quota proportionnel sont effectivement respectées en espérance mathématique avec un tirage aléatoire simple, il nous semble que le modèle d'échantillonnage devient *a priori* plus fragile quand les quotas s'éloignent significativement de quotas proportionnels, cela parce que le second temps de ce modèle est plus difficile à accepter. Si le premier temps du modèle est probablement la seule façon de construire une base théorique pour traiter les quotas, on pourrait en revanche – ce serait un autre exercice – chercher à maximiser la densité des $n_{i,j}$ sous les contraintes de quota pour obtenir les $\hat{P}_{i,j}$.

Jean-Claude Deville a proposé une expression de variance pour $\hat{\bar{Y}}_{\text{quota}}$, l'aléa de sondage étant ici le seul aléa qui génère de la variabilité. Il commence par décomposer les vraies moyennes par cellule $\bar{Y}_{i,j}$ selon $\bar{Y}_{i,j} = A_i + B_j + E_{i,j}$ en imposant les contraintes identifiantes $B_J = 0,$

$$\sum_{j=1}^{J} N_{i,j} E_{i,j} = 0, \quad \text{pour tout} \quad i = 1,\ldots,I$$

et

$$\sum_{i=1}^{I} N_{i,j} E_{i,j} = 0, \quad \text{pour tout} \quad j = 1,\ldots,J-1.$$

Puis, il introduit les coefficients a_i^0 et b_j^0 assurant les égalités

$$E_p\left\{\frac{n_{i,j}}{n}(a_i + b_j)^{-1}\right\} = \frac{N_{i,j}}{N}(a_i^0 + b_j^0)^{-1}.$$

En effet, $\frac{n_{i,j}}{n}(a_i + b_j)^{-1}$ est un estimateur de $P_{i,j} = \frac{N_{i,j}}{N}$ mais il n'est pas sans biais de $P_{i,j}$ dans le cas général, et les coefficients a_i^0 et b_j^0 vérifient les équations du système (S) considéré en espérance mathématique, soit

$$\sum_{j=1}^{J} N_{i,j}(a_i^0 + b_j^0)^{-1} = N_{i,.}, \quad \text{pour tout} \quad i = 1,\ldots,I,$$

$$\sum_{i=1}^{I} N_{i,j}(a_i^0 + b_j^0)^{-1} = N_{.,j}, \quad \text{pour tout} \quad j = 1,\ldots,J-1.$$

Si les quotas sont proportionnels, alors $a_i = 1$ et $b_j = 0$ et compte tenu du modèle multinomial, on a également $a_i^0 = 1$ et $b_j^0 = 0$ (ces valeurs sont des solutions du système non linéaire précédent dans tous les cas de figure, mais ce ne sont pas les valeurs adéquates pour des quotas non proportionnels). En posant

$$S_{i,j}^2 = \frac{1}{N_{i,j}} \sum_{k \in (i,j)} (y_k - \bar{Y}_{i,j})^2,$$

on en tire une expression de variance :

$$V(\hat{\bar{Y}}_{\text{quota}}) = \frac{1}{n} \sum_{(i,j)} \frac{N_{i,j}}{N} \{E_{i,j}^2 + (a_i^0 + b_j^0)^{-1} S_{i,j}^2\}.$$

La stratégie optimale n'est pas celle des quotas proportionnels, mais celle qui « gonfle » les quotas $n_{i.}$ et $n_{.j}$ pour les modalités i et j correspondant aux fortes dispersions $S_{i,j}^2$. Cette règle, très intuitive, n'apparait pas clairement dans son principe si on s'en tient à ce calcul de variance. En revanche, elle devient évidente quand on raisonne conditionnellement aux tailles $n_{i,j}$.

Lorsque les quotas sont proportionnels, on a donc

$$V(\hat{\bar{Y}}_{\text{quota}}) = \frac{1}{n} \sum_{(i,j)} \frac{N_{i,j}}{N} (E_{i,j}^2 + S_{i,j}^2).$$

On a clairement intérêt à avoir $E_{i,j} = 0$ pour tout (i, j) donc un modèle additif. L'article propose par ailleurs un estimateur de variance à faible biais qui n'est pas compliqué lorsque les quotas sont proportionnels parce qu'il s'obtient dans ce cas à partir d'un calcul classique de résidus dans une certaine régression linéaire standard.

6.3 Des modèles portant sur la variable d'intérêt

Il s'agit dans toute cette partie d'adopter un point de vue radicalement différent du précédent, puisque cette fois le modèle porte sur la variable d'intérêt Y. Cette variable est considérée comme aléatoire selon l'approche basée sur le modèle développée par Royall (1970, 1976a, b, 1988) (voir aussi Valliant, Dorfman et Royall, 2000; Chambers et Clark, 2012). Toutes les déclinaisons de modèle envisagées dans l'article s'inscrivent dans le cadre général du modèle linéaire, de type $\mathbf{Y} = \mathbf{XB} + \mathbf{u}$ où $\mathbf{B}$ est un vecteur de coefficients réels inconnus, $\mathbf{X}$ une matrice dont les colonnes constituent des variables explicatives et $\mathbf{u}$ est un vecteur de résidus d'espérance nulle et de variance sous le modèle $\mathbf{V}$ inconnue.

Dans ce contexte, un vecteur de valeurs y_k où k décrit la population U est composé de deux vecteurs, respectivement $\mathbf{Y}_s$ de taille n regroupant les valeurs y_k observées – donc celles où $k \in s$ – et $\mathbf{Y}_r$ de taille $N - n$ regroupant toutes les valeurs y_k non observées. Pour définir un estimateur optimum du vrai total $t_y = \sum_{k \in U} y_k$, il est naturel de chercher à minimiser l'erreur quadratique $\mathrm{E}_m(\hat{t} - t_y)^2$ où $\mathrm{E}_m(\cdot)$ désigne l'espérance mathématique lorsque l'aléa est celui qui affecte les valeurs y_k (qui n'a donc rien à voir avec l'aléa de sondage). On impose par ailleurs une forme d'absence de biais qui se traduit par l'égalité $\mathrm{E}_m(\hat{t} - t_y) = 0$. Enfin, on cherche un estimateur simple, donc linéaire, c'est-à-dire de forme $\hat{t} = \mathbf{g}_s^\top \mathbf{Y}_s$ où $\mathbf{g}_s$ est un vecteur (restant à trouver) de taille n. La solution du problème conduit à l'optimum

$$\hat{t}_{y,\text{opti}} = \mathbf{1}_s^\top \mathbf{Y}_s + \mathbf{1}_r^\top \left\{ \mathbf{X}_r \hat{\mathbf{B}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \hat{\mathbf{B}}) \right\},$$

où $\mathbf{1}_s$ (respectivement $\mathbf{1}_r$) est un vecteur de taille n (respectivement $N - n$) composé de 1, les termes $\mathbf{X}_s$, $\mathbf{X}_r$, $\mathbf{V}_{ss}$ et $\mathbf{V}_{rs}$ représentant les sous matrices de $\mathbf{X}$ et $\mathbf{V}$ formées par les lignes/colonnes associées aux

ensembles d'indices s et/ou r. Le coefficient estimé $\hat{\mathbf{B}}$ est celui des moindres carrés généraux, soit $\hat{\mathbf{B}} = (\mathbf{X}_s^\top \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} (\mathbf{X}_s^\top \mathbf{V}_{ss}^{-1} \mathbf{Y}_s)$. On va se placer *a priori* dans le cas où $\mathbf{V}_{rs} = 0$, hypothèse simplificatrice facile à accepter sauf si le tirage fait intervenir des grappes. De ce fait, l'estimateur optimum se simplifie beaucoup puisque

$$\hat{t}_{y,\text{opti}} = \mathbf{1}_s^\top \mathbf{Y}_s + \mathbf{1}_r^\top (\mathbf{X}_r \hat{\mathbf{B}}) = \sum_{k \in s} y_k + \sum_{k \in r} \hat{y}_k,$$

où $\hat{t}_{yk}$ est la k^{e} coordonnée du vecteur $\mathbf{X}_r \hat{\mathbf{B}}$, autrement dit le prédicteur optimum de la valeur inconnue y_k. C'est cette expression qui va servir par la suite pour exprimer les estimateurs issus de la méthode des quotas. Notons que si la constante fait partie des régresseurs (ce qui est extrêmement courant), on a $\sum_{k \in s} y_k = \sum_{k \in s} \hat{y}_k$, si bien que $\hat{t}_{y,\text{opti}} = \sum_{k \in U} \hat{y}_k$.

L'estimateur $\hat{t}_{y,\text{opti}}$ est par construction sans biais au sens où $\mathrm{E}_m(\hat{t}_{y,\text{opti}} - t_y) = 0$. L'appréciation de sa qualité fait aussi appel à sa variance, définie par $\mathrm{E}_p \mathrm{E}_m (\hat{t}_{y,\text{opti}} - t_y)^2$ où $\mathrm{E}_p(\cdot)$ désigne l'espérance mathématique par rapport à l'aléa de sondage. C'est aussi $\mathrm{E}_m \mathrm{E}_p (\hat{t}_{y,\text{opti}} - t_y)^2$.

L'association d'un aléa de sondage et d'un aléa de modèle n'est pas raisonnablement traitable si le modèle portant sur y_k change lorsqu'on possède une information sur l'appartenance – ou non – de l'individu k à l'échantillon tiré : sinon, tout calcul deviendrait ingérable, et au demeurant on ne serait pas même en mesure de formaliser de manière crédible la dépendance du modèle à l'échantillon, dont on rappelle qu'il est aléatoire. Puisqu'il n'est pas souhaitable (et pas possible pratiquement) de chercher à raffiner les hypothèses au-delà d'un certain degré, on considère que la loi de sélection de s et celle qui génère Y sont indépendantes. On parle de modèle non informatif. Dans ces conditions, le modèle posé sur y_k s'applique aveuglement à tout individu k sans qu'il y ait à se préoccuper de savoir si $k \in s$ ou si $k \notin s$. Ce parti pris est essentiel pour pouvoir mener les calculs de variance mais il est hélas contestable : typiquement dans le cas des sondages empiriques on peut douter, en tout cas lorsque l'opération n'est pas menée sous un contrôle très serré de la pratique des enquêteurs, qu'il n'y ait pas quelque relation entre les valeurs de Y et le fait d'appartenir ou non à l'échantillon. C'est d'ailleurs ce qui crée le principal risque de biais des estimateurs et ce qui alimente la critique traditionnelle des échantillonnages empiriques. Mais passons sur ce risque, acceptons-le, et résumons, dans ces conditions, la théorie exposée dans l'article de Jean-Claude Deville (1991).

Le cas des quotas fondés sur une seule variable qualitative et celui des quotas croisés relèvent du même modèle linéaire, qui est très simple : notant i la modalité de la variable de quota et k l'identifiant de l'individu, on pose $y_{i,k} = m_i + u_{i,k}$ avec $\mathrm{E}_m(u_{i,k}) = 0$ et $\mathrm{E}_m(u_{i,k}^2) = \sigma_i^2$. Cela traduit la situation naturelle espérée, celle où la variable qualitative (ou le croisement des deux variables) explique bien la variable d'intérêt. La technique de sondage fait que les tailles d'échantillon par modalité n_i sont indépendantes de l'échantillon tiré s. On vérifie que $\hat{y}_{i,k} = \hat{m}_i = \overline{y}_i$, moyenne simple des $y_{i,k}$ calculée sur les individus de l'échantillon vérifiant la modalité i. Dans ce cas, on vérifie facilement que l'estimateur optimum sans biais s'écrit

$$\hat{t}_{y,\text{opti}} = \sum_{i=1}^{I} N_i \bar{y}_i.$$

C'est exactement l'expression de l'estimateur post stratifié classique. On peut montrer que

$$\mathrm{E}_m \mathrm{E}_p (\hat{t}_{y,\text{opti}} - t_y)^2 = \sum_{i=1}^{I} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{\sigma_i^2}{n_i}$$

en exploitant le fait que les n_i sont indépendants de s, et obtenir ensuite une estimation sans biais de cette variance.

Considérons maintenant le cas des quotas marginaux, dans un contexte de modèle additif, c'est-à-dire, pour un individu k vérifiant respectivement les modalités i et j des deux variables de quota : $y_{i,j,k} = \alpha_i + \beta_j + u_{i,j,k}$ avec $\mathrm{E}_m(u_{i,j,k}) = 0$ et $\mathrm{E}_m(u_{i,j,k}^2) = \sigma_i^2 + \gamma_j^2$. Les résidus sont par ailleurs considérés comme deux à deux indépendants. L'estimateur optimum sans biais devient, en toute généralité

$$\hat{t}_{y,\text{opti}} = \sum_{i,j} n_{i,j} \bar{y}_{i,j} + \sum_{i,j} (N_{i,j} - n_{i,j})(\hat{\alpha}_i + \hat{\beta}_j).$$

Les estimateurs de α_i et de β_j étant compliqués à obtenir dans le cadre général ici présenté, Jean-Claude Deville propose d'utiliser les estimateurs des moindres carrés associés à un modèle ordinaire. Dans ce contexte, après avoir écrit les équations normales, et obtenu ainsi facilement les $\hat{\alpha}_i$ et $\hat{\beta}_j$, un peu d'algèbre amène à

$$\hat{t}_{y,\text{opti}} = \sum_i N_{i.} \hat{\alpha}_i + \sum_j N_{.j} \hat{\beta}_j.$$

Dans le cas des quotas proportionnels, on vérifie que $\hat{t}_{y,\text{opti}} = N\bar{y}$: c'est l'estimateur que l'on utilise classiquement dans le cadre des enquêtes par quotas, dont l'extrême simplicité constitue d'ailleurs un argument essentiel d'usage. Dans le cas des quotas non proportionnels, on peut toujours exprimer analytiquement $\hat{t}_{y,\text{opti}}$, en particulier comme une combinaison linéaire des estimateurs simples par cellule $\bar{y}_{i,j}$, soit $\hat{t}_{y,\text{opti}} = \sum_{i,j} \hat{N}_{i,j} \bar{y}_{i,j}$, mais les expressions des $\hat{N}_{i,j}$ sont complexes.

Lorsque les quotas sont proportionnels, la vraie variance $V(\hat{t}_{y,\text{opti}})$ peut s'exprimer en fonction des $N_{i.}$, $N_{.j}$, σ_i^2 et γ_j^2, et on peut également obtenir un estimateur de variance en formant

$$\hat{\mathrm{V}}(\hat{t}_{y,\text{opti}}) = N^2 \frac{1 - \frac{n}{N}}{n} \sum_{i,j} \frac{n_{i,j}}{n} s_{i,j}^2,$$

où $s_{i,j}^2$ est la dispersion standard des $y_{i,j,k}$ collectés dans la cellule (i,j). Il est sans biais en ce sens où $\mathrm{E}_p \mathrm{E}_m \hat{\mathrm{V}}(\hat{t}_{y,\text{opti}}) = \mathrm{V}(\hat{t}_{y,\text{opti}})$. La faisabilité d'un calcul de variance et de son estimation sans biais est intimement liée au caractère proportionnel des quotas : sans cette propriété, le développement de l'expression de la variance n'est plus possible parce que l'estimateur $\hat{t}_{y,\text{opti}}$ dépend des tailles d'échantillon par cellule $n_{i,j}$ et que celles-ci dépendent elles-mêmes de l'échantillon tiré… selon une probabilité que l'on ne connaît pas, par définition d'un tirage empirique.

L'article aborde enfin le cas, plus général, des quotas marginaux, dans un contexte où le modèle perd son caractère additif et devient $y_{i,j,k} = \alpha_i + \beta_j + \gamma_{i,j} + u_{i,j,k}$. L'estimateur optimum est $\hat{t}_{y,\text{opti}} = \sum_{i,j} N_{i,j}\, \overline{y}_{i,j}$ mais passé ce stade on se trouve dans un contexte où non seulement on ne connaît pas $N_{i,j}$ mais où il ne semble pas même possible d'en produire un estimateur « naturel ». Si les quotas sont proportionnels, la pratique – et manifestement seule issue possible – consiste à se rabattre sur l'expression simple et standard $\hat{t} = N\overline{y}$. Le prix à payer est celui d'un biais, qui vaut

$$\mathrm{E}_p\mathrm{E}_m(N\overline{y} - t_y) = \mathrm{E}_p\left\{\sum_{i,j}\left(N\frac{n_{i,j}}{n} - N_{i,j}\right)\gamma_{i,j}\right\}.$$

Il n'a aucune raison d'être nul, mais quand les consignes sont bien conçues et que la collecte les respecte, on peut espérer que les ratios $n_{i,j}/n$ et $N_{i,j}/N$ seront numériquement proches, et donc le biais faible.

6.4 Conclusion opérationnelle

La conclusion que l'on peut produire sur le traitement des enquêtes par quotas est la suivante. Si on dispose d'une seule variable de quotas ou si on pratique les quotas croisés (ce qui est le mieux!) en utilisant plusieurs variables de quotas, il n'y a aucune difficulté à estimer sans biais puis à calculer une variance. S'agissant de quotas marginaux, avec un modèle portant sur l'échantillonnage, on obtient un estimateur en toute circonstance. Il est asymptotiquement sans biais et sa variance peut toujours être estimée. Avec un modèle portant sur la variable d'intérêt, on ne sait pas gérer les modèles non additifs. Si le modèle est additif, on peut toujours produire un estimateur sans biais, mais si les quotas ne sont pas proportionnels alors on ne sait pas estimer de variance.

7. Conclusion

Ce qui précède ne couvre pas la totalité des développements originaux réalisés par Jean-Claude Deville dans le domaine de la statistique d'enquête. Il a en effet écrit de nombreux articles et effectués dans des colloques un grand nombre de présentations sur d'autres aspects moins porteurs. Cela étant, ses principaux travaux ont eux-mêmes connu des fortunes diverses. Le calage a certainement constitué l'avancée la plus marquante, la mieux connue et la plus utilisée à l'échelle mondiale, à tel point qu'aujourd'hui il n'existe plus d'enquête par sondage qui ne soit pas calée sur quelques données externes, au moins dans l'univers de la statistique publique. Le partage des poids relève de conditions d'enquête assez particulières mais il est néanmoins très utilisé, et probablement l'est-il surtout lorsqu'on met en œuvre des enquêtes répétées dans le temps et plus particulièrement sous forme d'échantillonnage rotatif. Le développement analytique de l'expression de la variance pour des estimateurs complexes trouve régulièrement des applications mais s'appuie sur une théorie assez compliquée et reste sérieusement concurrencé par les techniques de réplication d'échantillon. L'échantillonnage équilibré a également connu un certain succès, mais son

utilisation en statistique officielle se développe plus lentement que les méthodes de calage, probablement parce la modification d'un plan d'échantillonnage est une décision beaucoup plus lourde de conséquence que le changement d'une procédure d'estimation. Dix ans après le développement de la méthode, Tillé (2011) énumérait déjà une liste d'applications. L'intérêt pour l'échantillonnage équilibré a depuis continué à se développer. En France, l'Insee fait maintenant un très large usage des plans équilibrés. Quant à la théorie des quotas, il s'agit surtout d'une opération d'éclaircissement qui a permis de démystifier quelque peu cette technique que la statistique publique regarde toujours avec méfiance. Jean-Claude Deville aura décidément touché à tout, et avec brio, ce que la communauté scientifique a bien volontiers reconnu en lui attribuant le prestigieux prix Waksberg en 2018, trois années seulement avant sa disparition.

Bibliographie

Antal, E., Langel, M. et Tillé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. Conférence invitée au 58e Congrès de l'International Statistical Institute.

Bellhouse, D.R. (1988). A brief history of random sampling methods. *Handbook of Statistics Volume 6: Sampling*, (Éds., P.R. Krishnaiah et C.R. Rao), New York, Amsterdam. Elsevier/North-Holland, 1-14.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Bethlehem, J.G., et Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 34-42.

Binder, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases : Une approche de type « recette ». *Techniques d'enquête*, 22, 1, 17-22. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1996001/article/14389-fra.pdf.

Binder, D.A., et Kovačević, M.S. (1995). Estimation de l'inégalité du revenu d'après les données d'enquête : application de la méthode des équations d'estimation. *Techniques d'enquête*, 21, 2, 151-159. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1995002/article/14396-fra.pdf.

Boistard, H., Lopuhaä, H.P. et Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electronic Journal of Statistics*, 6, 1967-1983.

Bousabaa, A., Lieber, J. et Sirolli, R. (1999). *La Macro Cube*. Rapport technique, Rennes: ENSAI.

Brewer, K. (2013). Trois controverses dans l'histoire de l'échantillonnage. *Techniques d'enquête*, 39, 2, 275-289. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2013002/article/11883-fra.pdf.

Chambers, R.L., et Clark, R.G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford: Oxford University Press.

Chang, T., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Chauvet, G. (2009). Échantillonnage équilibré stratifié. *Techniques d'enquête*, 35, 1, 123-127. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2009001/article/10888-fra.pdf.

Chauvet, G., Bonnéry, D. et Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141(2), 984-994.

Chauvet, G., Deville, J.-C. et Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98, 459-471.

Chauvet, G., Haziza, D. et Lesage, É. (2015). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*, 25, 313-334.

Chauvet, G., et Tillé, Y. (2006a). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21, 9-31.

Chauvet, G., et Tillé, Y. (2006b). Fastcube SAS-IML Macro. Université de Neuchâtel.

Chopin, N., et Ducrocq, G. (2021). Fast compression of MCMC output. *Entropy*, 23(8).

Choudhry, G.H., et Singh, M.P. (1979). Échantillonnage avec probabilités inégales et sans remise - une méthode réjective. *Techniques d'enquête*, 5, 2, 1-14. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1979002/article/54834-fra.pdf.

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, 18(5), 353-362.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are know. *Annals of Mathematical Statistics*, 11, 427-444.

Demnati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 1, 17-27. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2004001/article/6991-fra.pdf.

Demnati, A., et Rao, J.N.K. (2010). Estimateurs de variance par linéarisation pour les paramètres de modèles à partir de données d'enquêtes complexes. *Techniques d'enquête*, 36, 2, 211-220. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010002/article/11381-fra.pdf.

Devaud, D., et Tillé, Y. (2019). Deville and Särndal's calibration: Revisiting a 25 years old successful optimization problem. *TEST*, 4, 1033-1065.

Deville, J.-C. (nda). 15ème round. cette fois y'a vraiment qêqchose. Note interne manuscrite, Paris: Insee.

Deville, J.-C. (ndb). Comparaison des plans à probabilités inégales avec et sans remise. Note interne manuscrite, Paris: Insee.

Deville, J.-C. (ndc). Échantillonnage à entropie max (rédaction rapide). Note interne manuscrite, Paris: Insee.

Deville, J.-C. (1989). Une théorie simplifiée des sondages. *Les Ménages : Mélanges en L'honneur de Jacques Desabie*, (Éds., P. L'Hardy et C. Thélot), Paris: Insee, 191-214.

Deville, J.-C. (1991). Une théorie des enquêtes par quotas. *Techniques d'enquête*, 17, 2, 177-195. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1991002/article/14504-fra.pdf.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro. Statistics Sweden, 1-18.

Deville, J.-C. (1998a). La correction de la nonréponse par calage ou par échantillonnage équilibré. *Actes du Colloque de la Société Statistique du Canada,* Sherbrooke, Canada.

Deville, J.-C. (1998b). Les enquêtes par panel : en quoi différent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d'une autre. *Actes des Journées de Méthodologie Statistique,* INSEE Méthodes, Paris: Insee, No. 84-85-86, 63-82.

Deville, J.-C. (1998c). Une nouvelle (encore une!) méthode de tirage à probabilités inégales. Document de travail nº 9804, Méthodologie Statistique, Paris: Insee.

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. (2000a). Generalized calibration and application to weighting for non-response. *CompStat, Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*, New York: Springer, 65-76.

Deville, J.-C. (2000b). Note sur l'algorithme de Chen, Dempster et Liu. Note interne manuscrite, Rennes: CREST-ENSAI.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Paris: Insee-Méthodes, 3-20.

Deville, J.-C. (2014). Échantillonnage équilibré exact poissonnien. *8ème Colloque Francophone sur les Sondages*, Université de Bourgogne, Dijon, 1-6.

Deville, J.-C. (2015). Quelques éléments de géométrie et d'algèbre pour comprendre la nature d'un échantillonnage équilibré. *Actes des Journées de Méthodologie Statistique*, Paris: Insee-Méthodes, 1-8.

Deville, J.-C., et Dupont, F. (1993). Non-réponse : principes et méthodes. *Actes des Journées de Méthodologie Statistique*, Paris: Insee-Méthodes, 53-69.

Deville, J.-C., et Grosbras, J.-M. (1987). Algorithmes de tirage. *Les Sondages*, (Éds., J.-J. Droesbeke, B. Fichet et P. Tassi), Paris: Economica, 209-233.

Deville, J.-C., Grosbras, J.-M. et Roth, N. (1988). Efficient sampling algorithms and balanced sample. *CompStat, Proceedings in Computational Statistics 8th Symposium held in Copenhagen*, Heidelberg: Physica Verlag, 255-266.

Deville, J.-C., et Jacod, M. (1996). Replacing the traditional French census by a large scale continuous population survey. *Annual Research Conference*. US Bureau of Census.

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 Maumy-Bertrand, M. (2006). Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme. *Techniques d'enquête*, 32, 2, 197-206. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9552-fra.pdf.

Deville, J.-C., et Qualité, L. (2005). Échantillonnage multidimensionnel (de plusieurs échantillons à la fois) à entropie maximum : définition, propriétés, algorithmes et programmes. *Actes des Journées de Méthodologie Statistique*, Paris: Insee-Méthodes, 1-8.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

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

Deville, J.-C., et Tillé, Y. (2000a). Échantillonnage équilibré par la méthode du cube, variance et estimation de variance. *Actes des Journées de Méthodologie Statistique*, Paris: Insee-Méthodes, 15-35.

Deville, J.-C., et Tillé, Y. (2000b). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, 86, 215-227.

Deville, J.-C., et Tillé, Y. (2001). Échantillonnage équilibré par la méthode du cube, variance et estimation de variance. *Enquêtes, Modèles et Applications*, (Éds., J.-J. Droesbeke et L. Lebart), Paris: Dunod, 444-362.

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

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Dupačová, J. (1979). A note on rejective sampling. *Contribution to Statistics (Jaroslav Hájek Memorial Volume)*, Academia Prague, 71-78.

Dupont, F. (1993). Calage et redressement de la non-réponse totale : validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. *Actes des Journées de Méthodologie Statistique,* 15 et 16 décembre 1993, Insee-Méthodes No 56-57-58, Complément, 9-42.

Durr, J.-M., et Dumais, J. (2002). La rénovation du recensement français. *Techniques d'enquête*, 28, 1, 47-53. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2002001/article/6414-fra.pdf.

Ernst, L.R. (1989). Weighting issues for longitudinal household and family estimates. *Panel Surveys*, (Éds., D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh), New York: John Wiley & Sons, Inc., 135-159.

Ernst, L.R., Hubble, D.L. et Judkins, D.R. (1984). Longitudinal family and household estimation in sipp. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 682-687.

Estevao, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in twophase sampling. *Journal of Official Statistics*, 18, 233-255.

Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.

Eustache, E., Vallée, A.-A. et Tillé, Y. (2022). Balanced donor imputation handling Swiss cheese nonresponse. *Statistica Sinica*, accepté.

Falorsi, P., Piersante, A. et Bako, B. (2016). Indirect sampling, a way to overcome the weakness of the lists in agricultural survey. *Proceedings ICAS VII: Seventh International Conference on Agricultural Statistics*, Rome, 24 au 26 octobre 2016.

Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 454-469.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A., Legg, J.C. et Li, Y. (2017). Bootstrap variance estimation for rejective sampling. *Journal of the American Statistical Association*, 112(520), 1562-1570.

Gini, C., et Galvani, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). *Annali di Statistica*, Series 6, 4, 1-107.

Graf, É., et Tillé, Y. (2014). Estimation de variance par linéarisation pour des indices de pauvreté et d'exclusion sociale. *Techniques d'enquête*, 40, 1, 69-88. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2014001/article/14000-fra.pdf.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. *Compositional Data Analysis: Theory and Applications*, (Éds., V. Pawlowsky-Glahn et A. Buccianti), Chichester: Wiley, 114-127.

Grafström, A., et Lisic, J. (2019). *BalancedSampling: Balanced and Spatially Balanced Sampling*. R package version 1.5.5.

Grafström, A., Lundström, N.L.P. et Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514-520.

Grafström, A., et Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 14(2), 120-131.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J. et Stahel, W.A. (1985). *Robust Statistics: The Approach Based on the Influence Function*. New York: John Wiley & Sons, Inc.

Hasler, C., et Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74, 81-94.

Hasler, C., et Tillé, Y. (2016). Balanced k -nearest neighbor imputation. *Statistics*, 105, 11-23.

Haziza, D., et Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32(2), 206-226.

Haziza, D,. et Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.

Holt, D., et Smith, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*, A142: Part 1, 33-46.

Huang, E.T., et Fuller, W.A. (1978). Non-negative regression estimation for sample survey data. *Proceedings of the Social Statistics Section,* American Statistical Association, 300-305.

Huang, H. (1984). Obtaining cross-sectional estimates from a longitudinal survey: Experiences of the income survey development program. *Proceedings of the Social Statistics Section,* American Statistical Association, 670-675.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77, 89-96.

Jauslin, R., Eustache, E., Panahbehagh, B. et Tillé, Y. (2021). *StratifiedSampling: Different Methods for Stratified Sampling*. R Foundation for Statistical Computing, Vienne, Autriche. R package version 0.3.0.

Jauslin, R., Eustache, E. et Tillé, Y. (2021). Enhanced cube implementation for highly stratified population. *Japanese Journal of Statistics and Data Science*, 4, 783-795.

Jauslin, R., et Tillé, Y. (2020a). Spatial spread sampling using weakly associated vectors. *Journal of Agricultural, Biological and Environmental Statistics*, 25(3), 431-451.

Jauslin, R., et Tillé, Y. (2020b). *WaveSampling: Weakly Associated Vectors Sampling*. R package version 0.1.1 http://CRAN.R-project.org/package=WaveSamplinghttp://CRAN.R-project.org/package=WaveSampling.

Judkins, D., Hubble, D., Dorsch, J., McMillen, D. et Ernst, L. (1984). Weighting of persons for sipp longitudinal tabulations. *Proceedings of the Social Statistics Section,* American Statistical Association, 676-687.

Kalton, G., et Brick, J. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 1, 37-49. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1995001/article/14412-fra.pdf.

Kiær, A.N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9, 176-183.

Kiær, A.N. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 180-185.

Kiær, A.N. (1903). Sur les méthodes représentatives ou typologiques. *Bulletin de l'Institut International de Statistique*, 13, 66-78.

Kiær, A.N. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 119-134.

Kiesl, H. (2010). Selecting kindergarten children by three stage indirect sampling. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 2730-2738.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2006002/article/9547-fra.pdf.

Kott, P.S., et Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491), 1265-1275.

Langel, M., et Tillé, Y. (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. *Metron*, 69, 45-65.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 1, 27-35. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1995001/article/14413-fra.pdf.

Lavallée, P. (2001). La méthode généralisée du partage des poids et le calage sur marges. *Enquêtes, Modèles et Applications*, (Éds., J.-J. Droesbeke et L. Lebart), Paris: Dunod, 396-403.

Lavallée, P. (2002). *Le Sondage Indirect ou la Méthode Généralisée du Partage des Poids*. (Éd. de l'Université de Bruxelles, Paris: Ellipses.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Legg, J.C., et Yu, C.L. (2010). Comparaison de méthodes de restriction de l'ensemble d'échantillons. *Techniques d'enquête*, 36, 1, 75-87. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010001/article/11249-fra.pdf.

Lemel, Y. (1976). Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondages. *Annales de l'Insee*, 22-23, 273-281.

Lesage, É., Haziza, D. et D'Haultfoeuille, X. (2019). A cautionary tale on instrumental calibration for the treatment of nonignorable unit nonresponse in surveys. *Journal of the American Statistical Association*, 114, 906-915.

Leuenberger, M., Eustache, E., Jauslin, R. et Tillé, Y. (2022). Balancing a sample almost perfectly. *Statistics and Probablity Letters*, 180, 109229.

Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

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 traffc estimation. *The Annals of Applied Statistics*, 17(1), 838-859.

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

Nedyalkova, D., et Tillé, Y. (2012). Bias robustness and efficiency in model-based inference. *Statistica Sinica*, 22, 777-794.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. *Biometrika*, 63, 605-614.

Royall, R.M. (1976b). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R.M. (1988). The prediction approach to sampling theory. *Handbook of Statistics, Sampling*, (Éds., P.R. Krishnaiah et C.R. Rao), Amsterdam: Elsevier, Volume 6, 399-413.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Särndal, C.-E. (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, 33, 2, 113-135. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2007002/article/10488-fra.pdf.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Sautory, O., et Le Guennec, J. (2003). La macro CALMAR2 : redressement d'un échantillon par calage sur marges – documentation de l'utilisateur. Rapport technique, Paris: Insee.

Stephan, F.F. (1942). An iterative method of adjusting sample frequency data tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.

Thionet, P. (1953). *La Théorie des Sondages*. Institut National de la Statistique et des Études Économiques, Études théoriques vol. 5, Paris: Imprimerie nationale.

Thomsen, I. (1978). A second note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift*, 16, 278-285.

Tillé, Y. (2011). Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation. *Techniques d'enquête*, 37, 2, 233-246. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2011002/article/11609-fra.pdf.

Tillé, Y. (2016). The legacy of Corrado Gini in survey sampling and inequality theory. *Metron*, 74(2), 167-174.

Tillé, Y., et Matei, A. (2021). *Sampling: Survey Sampling*. R package version 2.9.

Vallée, A.-A., et Tillé, Y. (2019). Linearization for variance estimation by means of sampling indicators: Application to nonresponse. *Revue Internationale de Statistique*, 87(2), 347-367.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

Wu, C., et Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Londres: Charles Griffin.

Zhang, S., Han, P. et Wu, C. (2022). Calibration techniques encompassing survey sampling, missing data analysis and causal inference. *Revue Internationale de Statistique*, accepté pour publication.