Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

ESTIMATION ET COHÉRENCE PAR LA PONDÉRATION RÉPÉTÉE

(version abrégée)

R.H. Renssen, A.H. Kroese et A. J. Willeboordse[1]

RÉSUMÉ

À Statistics Netherlands, la conception et l'organisation du processus statistique évoluent rapidement, mues par la nécessité de produire des données plus cohérentes et par les pressions gouvernementales en vue de réduire le fardeau de réponse. Le nouveau processus de production cherche avant tout à intégrer toutes les données d'enquête et les données administratives dans un nombre limité de bases de microdonnées et à élaborer une stratégie d'estimation par base de microdonnées. La présente communication donne le coup d'envoi à cette nouvelle stratégie d'estimation. La stratégie proposée garantit que tous les tableaux d'estimations à *m* dimensions sont numériquement cohérents en ce qui a trait aux marges communes, même si les données de ces tableaux sont estimées d'après des enquêtes différentes. La stratégie se fonde toujours sur le principe du calage, mais ne requiert pas nécessairement un ensemble constant de poids par enquête. L'aspect pratique de la stratégie est illustré à l'aide d'un exemple fictif.

Mots-clés : Estimateurs par calage, Cohérence, Distribution de fréquences, Estimateurs par régression générale.

1. INTRODUCTION

Traditionnellement, les organismes statistiques organisent leurs méthodes de collecte, de traitement et de diffusion des données selon un modèle cloisonné; en effet, ils mènent bien des enquêtes différentes plus ou moins indépendamment les unes des autres, chaque enquête disposant de sa propre méthode de traitement. Cette méthode est insatisfaisante pour plusieurs raisons. Premièrement, il peut s'avérer impossible de comparer les données statistiques à cause du manque de cohérence entre les diverses enquêtes[2]. Deuxièmement, afin de limiter le fardeau de réponse, on doit interroger les répondants le moins souvent possible, alors que le contraire peut se produire lorsqu'on utilise le modèle cloisonné. Troisièmement, l'exactitude des estimations peut s'avérer inutilement faible si, dans la stratégie d'estimation, on utilise peu ou point les enregistrements supplémentaires ou les autres enquêtes.

Pour parer à ces inconvénients, Statistics Netherlands a décidé de transformer radicalement ses méthodes de production (voir, par exemple, Willeboordse (2000), Al et Bakker (2000) et Laan van der (2000)). Il s'agit d'intégrer toutes les sources primaires et secondaires de microdonnées en un nombre restreint de bases de microdonnées et d'élaborer une stratégie d'estimation efficace, de sorte que toutes les estimations présentées sous forme de tableaux à *m* dimensions soient numériquement cohérentes, c'est-à-dire qu'il ne puisse y avoir aucune contradiction lorsque l'on compare deux tableaux ou davantage, même en faisant abstraction de l'erreur d'échantillonnage.

À l'heure actuelle, Statistics Netherlands établit une distinction entre plusieurs bases de microdonnées, celle sur les personnes et celle sur les entreprises étant les plus importantes. De

---

[1] R.H. Renssen, A.H. Kroese et A.J. Willeboordse, Methods and Informatics Department, P.O. Box 4481, 6401 CZ, Heerlen (Pays-Bas). Les opinions exprimées dans la présente communication sont celles des auteurs et ne reflètent pas nécessairement les politiques de Statistics Netherlands. Les auteurs tiennent à remercier Peter Kooiman, Nico Nieuwenbroek et René Achenbach pour leur lecture attentive du document et leurs observations précieuses, ainsi que Jeroen Pannekoek pour sa précieuse collaboration.

[2] Willeboordse et Ypma (1996 et 1998) distinguent deux niveaux de cohérence, soit la « coordination des concepts », qui nécessite la coordination et l'uniformisation des variables et des classements et la « cohérence interne des données », qui nécessite l'harmonisation de la collecte et du traitement des données.

manière quelque peu simplifiée, la base de microdonnées sur les personnes se compose essentiellement de la base de données des administrations municipales, à laquelle s'ajoutent des enquêtes par sondage auprès des personnes et d'autres registres de personnes. De même, la base de microdonnées sur les entreprises se compose essentiellement du Registre général des entreprises, auquel s'ajoutent des enquêtes auprès des entreprises et d'autres registres d'entreprises. On considère ordinairement les bases de microdonnées comme des tableaux rectangulaires dont les rangées présentent des unités statistiques et les colonnes, des résultats liés à des variables. À l'évidence, seuls les résultats observés sont enregistrés. Les résultats non observés correspondent à des cases vides, étant entendu que des imperfections comme les erreurs de mesure et la non-réponse partielle ont déjà été corrigées au moyen d'une stratégie de contrôle et d'imputation.

La méthode traditionnelle de calcul d'estimations consiste à utiliser un seul ensemble de poids par enquête. Étant donné les probabilités d'inclusion de premier ordre, on peut obtenir un tel ensemble au moyen de techniques de calage, comme l'expliquent Deville et Särndal (1992). Lorsqu'on utilise un seul ensemble de poids par enquête, toutes les variables sont extrapolées de la même façon. Le principal avantage de cette méthode est qu'après avoir calculé l'ensemble de poids, on peut l'appliquer directement à n'importe quel ensemble de variables à étudier et obtenir des estimations numériquement cohérentes. Toutefois, cette méthode ne convient pas dans le cas d'une base de microdonnées, car on est en présence de plusieurs enquêtes (et registres). Si l'on obtient la cohérence numérique par enquête, bon nombre d'estimations sont numériquement incohérentes d'une enquête à l'autre, comme le montrent, par exemple, Kroese, Renssen et Trijssenaar (2000). Nous proposons une autre stratégie d'estimation, toujours fondée sur la pondération, mais pas nécessairement sur un seul ensemble de poids par enquête.

Avant de décrire notre stratégie d'estimation, nous présentons d'abord dans la section 2 quelques notions conceptuelles de base, dont une terminologie et une notation qui s'appliquent à la question. Nous proposons notamment le concept de *distribution générale de fréquences*, laquelle définit un cadre systématique qui devrait englober toutes les distributions de fréquences cibles potentielles. Ce cadre de travail est important, non seulement pour définir la notion de « cohérence numérique », mais aussi pour structurer la multitude de distributions cibles que les organismes statistiques doivent habituellement prendre en compte. La stratégie d'estimation est présentée dans la section 3. Elle comporte trois étapes. La première consiste à répartir la base de microdonnées en un certain nombre de sous-ensembles rectangulaires. La deuxième étape consiste à attribuer à chacun des sous-ensembles de microdonnées des poids de régression préliminaires permettant de calculer un grand nombre d'estimations. La troisième étape consiste à corriger les incohérences entre ces estimations. Dans la section 4, nous présentons un exemple numérique. Enfin, dans la section 5, nous abordons brièvement certains sujets nécessitant une recherche plus poussée.

2. LE CADRE CONCEPTUEL

Avant d'aborder la stratégie d'estimation proprement dite, il importe de bien comprendre la nature et la structure des données statistiques faisant l'objet d'une estimation. Dans la présente section, nous allons donc examiner d'abord les structures de données dans la mesure où elles s'appliquent aux *bases de microdonnées* et alimentent le processus d'estimation. Puis, nous expliquerons la structure des *bases de données agrégées* constituées à la suite du processus d'estimation. Nous définirons le cadre conceptuel en présentant une terminologie des concepts pertinents et une notation mathématique.

2.1 Terminologie

On suppose que chaque base de microdonnées couvre et définit une certaine population cible, soit l'ensemble complet d'*objets statistiques* qui instancient un certain type d'objet. Chaque objet forme une ligne dans la liste de la base de données. Dans le cas de la base de microdonnées sur les personnes, cet ensemble correspond à la base de données des administrations municipales; dans le cas de la base de microdonnées sur les entreprises, il correspond au Registre général des entreprises.

Chaque objet est décrit selon un ou plusieurs *attributs*, exprimés en termes de *valeurs* que prend l'objet à l'égard de diverses *variables*. Sur le plan opérationnel, ces variables sont définies en fonction d'une mesure spécifique d'échelle, ce qui fournit une méthode valide pour les mesurer. En ce qui concerne la mesure d'échelles, une distinction importante s'impose entre

- les variables *catégoriques* (ou *qualitatives*), mesurées sur une échelle *nominale* or *ordinale* dont les valeurs correspondent à des catégories qui sont souvent calculées à partir d'un classement;
- les variables *quantitatives*, mesurées sur une échelle de *ratios* ou d'*intervalles* dont les valeurs se présentent sous forme de nombres.

En outre, il importe de savoir qu'on peut définir les variables des deux types selon différents niveaux de généralisation. Nous proposons les qualificatifs *de premier ordre* pour une variable qui – étant donné les renseignements disponibles d'après l'observation – est définie au niveau le plus élémentaire (par exemple, le revenu en euros) et *de deuxième ordre* pour une variable dont on obtient les valeurs par généralisation à partir d'une variable de premier ordre (par exemple, le revenu par classe). On peut calculer les variables de deuxième ordre à partir de variables catégoriques ou quantitatives. Dans le premier cas, il s'agit de comprimer les catégories existantes pour les porter à un niveau supérieur dans une hiérarchie de classement; dans le deuxième, il s'agit de *créer* des catégories à partir de nombres. Sur le plan terminologique, les variables de deuxième ordre, calculées à partir de variables quantitatives de premier ordre, restent des variables « quantitatives ». D'ailleurs, il reste possible de les appliquer dans des opérations arithmétiques lorsqu'on se sert des renseignements fournis par les variables de premier ordre correspondantes.

En théorie, il est possible de calculer de nombreuses variables de deuxième ordre à partir de variables de premier ordre. Toutefois, afin d'empêcher la prolifération des paramètres d'une population finie, ce qui entraînerait de la confusion ou même de l'incohérence dans les données destinées à la publication, il convient de restreindre délibérément le nombre de catégories admissibles à titre de variables de deuxième ordre et de définir exactement ces variables dans la base de microdonnées, à la suite des variables de premier ordre. On peut obtenir un certain degré de restriction et de discipline en imposant la règle selon laquelle toutes les variables de deuxième ordre qui sont calculées à partir d'une certaine variable de premier ordre doivent être emboîtées. On dit qu'une variable *A* est *emboîtée* dans une variable *B* ou, plus brièvement, que *A* est emboîté dans *B* si chaque catégorie de *A* s'intègre à une seule catégorie de *B*. Concrètement, il s'ensuit de cette restriction que les variables de premier ordre peuvent s'accompagner uniquement d'une suite hiérarchique de variables de deuxième ordre.

Si l'on considère la structure des données au niveau des agrégats, une distinction supplémentaire (et non de rechange) s'impose, notamment à l'égard du *rôle* d'une variable dans le processus d'agrégation ou d'estimation, qui consiste à transformer les microdonnées en données agrégées :

- les variables *de classement*, soit les variables dont les valeurs (= catégories) servent à subdiviser le type d'objet de la base de microdonnées en sous-types et, par conséquent, à subdiviser la population en sous-populations ou classes;
- les variables *de quantification*, soit les variables dont les valeurs (= nombres) servent à compiler des totaux, des moyennes, etc., qui s'appliquent aux classes définies ci-dessus.

Soulignons que les variables de classement de la base de données agrégées ne sont pas seulement la contrepartie des variables catégoriques de la base de microdonnées. En effet, ce ne sont pas seulement les variables catégoriques, mais aussi les variables quantitatives de deuxième ordre qui peuvent servir de variables de classement. Soulignons en outre qu'une variable de classement peut constituer un *recoupement* de plusieurs variables (catégoriques ou quantitatives de deuxième ordre), par exemple l'âge et le sexe, ou plutôt l'âge *selon* le sexe. Il est donc logique d'établir une distinction entre les variables (de classement) simples et les variables (de classement) multiples, les premières correspondant à une seule variable catégorique ou quantitative de deuxième ordre et les deuxièmes, à plus d'une. On peut dire que les catégories de variables de classement multiples correspondent à des classements recoupés. Dans le reste du présent exposé, nous considérons

uniquement les variables de classement. Renssen et coll. (2001) traitent plus longuement des variables de quantification.

2.2 Notation

Supposons que *U* désigne une certaine population cible et $X^{(r)}$ une variable de classement simple comportant *p* classes. Pour simplifier la notation mathématique, nous représentons $X^{(r)}$ comme un vecteur *p* de variables auxiliaires; ainsi, nous définissons

$$\mathbf{x}_i^{(r)} = (x_{i1},...,x_{ip})^t \text{ où } x_{ij} = \begin{cases} 1 \text{ si le ième objet appartient à la classe j} \\ 0 \qquad\qquad\qquad \text{sinon} \end{cases}$$

Une *suite hiérarchique* de *k* variables de classement simples est désignée par $\mathbf{x}_i^{(1)},...,\mathbf{x}_i^{(k)}$. Ici, $\mathbf{x}_i^{(r)}$ est emboîté à l'intérieur de $\mathbf{x}_i^{(s)}$ si $s < r$. Selon cette notation, $\mathbf{x}_i^{(k)}$ contient le plus grand nombre de classes. Toujours pour simplifier la notation, nous définissons également $\mathbf{x}_i^{(0)} \equiv 1$. On peut considérer cette constante comme une variable de classement dégénérée qui correspond à une seule classe, soit la population entière. Si l'on distingue *G* variables (de classement) de premier ordre, il devient évident que le nombre *G* de suites hiérarchiques peut être défini comme suit :

$$\mathbf{x}_{1i}^{(1)},...,\mathbf{x}_{1i}^{(k_1)},\ \mathbf{x}_{2i}^{(1)},...,\mathbf{x}_{2i}^{(k_2)},\ldots,\mathbf{x}_{Ji}^{(1)},...,\mathbf{x}_{Gi}^{(k_G)}.$$

Supposons que $\mathbf{x}_{1i}^{(r_1)},\mathbf{x}_{2i}^{(r_2)},...,\mathbf{x}_{mi}^{(r_m)}$ désigne *m* variables de classement simples comportant respectivement $p_1, p_2,..., p_m$ classes, $0 \le r_j \le k_j$, $j = 1,...,m$. Le recoupement complet entre ces variables, qui est désigné par $\mathbf{x}_{1i}^{(r_1)} \otimes \mathbf{x}_{2i}^{(r_2)} \otimes ... \otimes \mathbf{x}_{mi}^{(r_m)}$, donne une variable de classement multiple comportant $p_1 \times p_2 \times ... \times p_m$ classes. La distribution de fréquences de la population de cette variable peut être énoncée comme suit :

$$\sum_{i \in U} \mathbf{x}_{1i}^{(r_1)} \otimes \mathbf{x}_{2i}^{(r_2)} \otimes ... \otimes \mathbf{x}_{mi}^{(r_m)}.$$

Soulignons qu'on obtient le total de la population en prenant toutes les $r_j = 0$; les distributions de fréquences unidimensionnelles en prenant toutes les $r_j = 0$ sauf une; les distributions de fréquences bidimensionnelles en prenant toutes les $r_j = 0$ sauf deux, etc.

Supposons que $\mathbf{x}_{1i}^{(k_1)} \otimes \mathbf{x}_{2i}^{(k_2)} \otimes ... \otimes \mathbf{x}_{Gi}^{(k_G)}$ constitue la variable de classement multiple qui répartit la population pour donner le nombre maximal de sous-populations ou de groupes disjoints. On peut alors considérer

$$\sum_{i \in U} \mathbf{x}_{1i}^{(k_1)} \otimes \mathbf{x}_{2i}^{(k_2)} \otimes ... \otimes \mathbf{x}_{Gi}^{(k_G)}. \qquad (1)$$

comme une *distribution générale de fréquences*. Elle constitue le cadre qui sert à définir la multitude de paramètres descriptifs simples d'une population finie que les organismes statistiques doivent habituellement prendre en compte. Seules les distributions de fréquences qu'on peut calculer à titre de distributions marginales à partir de (1), soit

$$\sum_{i \in U} \mathbf{x}_{1i}^{(r_1)} \otimes \mathbf{x}_{2i}^{(r_2)} \otimes ... \otimes \mathbf{x}_{gi}^{(r_g)}, \qquad (2)$$

où $0 \le r_g \le k_g$, $g = 1,...,G$, sont des candidats admissibles à la base de données agrégées. Toutes ces distributions sont manifestement liées aux distributions générales de fréquences par le biais d'un vecteur G $\mathbf{r} = (r_1,...,r_G)^t$. Le nombre de valeurs non nulles de ce vecteur représente la dimension de la distribution, tandis que ces valeurs non nulles correspondent aux niveaux de variables de classement simples en présence. D'un point de vue différent, $\mathbf{r}$ correspond à une partition de la population finie en un ensemble exclusif et exhaustif de sous-populations.

De par leur construction, deux distributions de fréquences $\mathbf{D}_1$ et $\mathbf{D}_2$ sont nécessairement liées entre elles; en effet, les totaux des cases ou les combinaisons de totaux des cases d'une distribution sont égaux aux totaux des cases ou aux combinaisons de totaux des cases de l'autre. Supposons que $\mathbf{D}_1$ soit caractérisée par $\mathbf{r}_1$ et $\mathbf{D}_2$ par $\mathbf{r_2}$; ces totaux communs définissent alors une distribution quantitative commune caractérisée par $\mathbf{s} = \min(\mathbf{r_1}, \mathbf{r_2})$, où le minimum est déterminé par les composantes. Compte tenu de l'exigence en matière de cohérence, cette distribution quantitative commune joue un rôle essentiel dans notre méthode d'estimation. Soulignons que si $\mathbf{D}_1$ et $\mathbf{D}_2$ sont définis par des ensembles de variables de classement simples sans chevauchement, elles comportent tout de même un élément commun, soit le total de la population, qui correspond à un vecteur $\mathbf{s}$ nul.

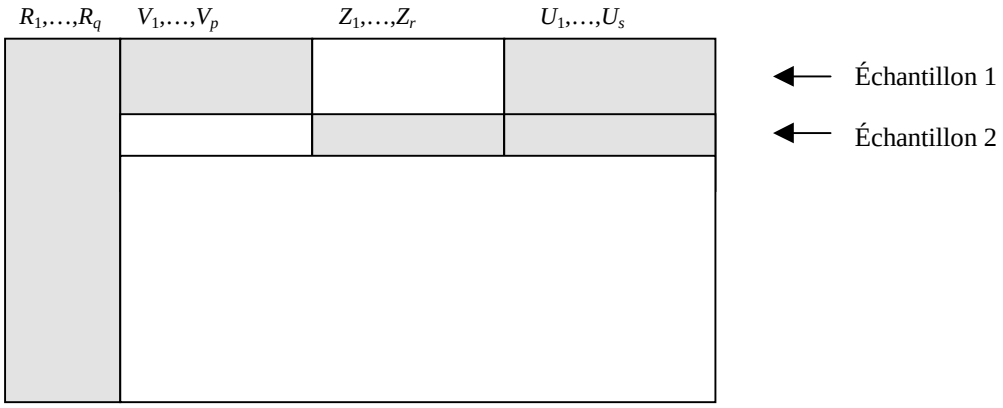
3. LA STRATÉGIE D'ESTIMATION

Dans la section précédente, nous avons présenté le cadre conceptuel qui devrait englober chaque distribution de fréquences envisagée en vue de la publication des données. Supposons que $\mathbf{D}$ désigne une telle fréquence. Pour être admissible à la publication, $\mathbf{D}$ doit habituellement répondre à trois conditions : $\mathbf{D}$ doit être pertinente, $\mathbf{D}$ doit être sûre, c'est-à-dire conforme aux règles de la non-divulgation, et $\hat{\mathbf{D}}$ doit être suffisamment exacte, c'est-à-dire que toutes les estimations des cases de $\mathbf{D}$ doivent présenter une erreur quadratique moyenne suffisamment faible. Toutes ces conditions sont bien connues et couramment imposées par la plupart des organismes statistiques; il est donc inutile de nous y attarder.

Afin de répondre aux attentes des utilisateurs en matière de comparabilité des produits statistiques, Statistics Netherlands impose aussi une quatrième condition : $\hat{\mathbf{D}}$ doit être cohérente à l'égard de toutes les distributions connexes déjà publiées (et stockées) auparavant, même si ces distributions ont été estimées à partir d'autres enquêtes. La stratégie d'estimation qui tient compte de cette quatrième condition est nouvelle et fera l'objet de la présente section. Elle est fondée sur l'estimateur par régression générale, quoiqu'elle aurait pu, de façon plus générale, se baser sur l'estimateur par calage. Elle comporte trois étapes : 1) construire des sous-ensembles rectangulaires de microdonnées à partir d'une certaine base de microdonnées, 2) attribuer à chaque sous-ensemble de microdonnées un ensemble de poids de régression selon un système de pondération, et 3) établir une estimation cohérente d'un ensemble de distributions quantitatives.

En guise d'illustration, prenons la version simplifiée représentée à la figure 1. Seules les surfaces ombrées renferment des observations. Les colonnes *R* correspondent à un enregistrement complet, auquel deux enquêtes par sondage sont appariées. Dans l'une de ces enquêtes, désignée S_1, on observe les variables *V* et *U*; dans l'autre, désignée S_2, on observe les variables *Z* et *U*. On observe donc les variables *U* dans les deux enquêtes par sondage. Nous remarquons que le registre correspond précisément à la population (finie) qui nous intéresse. Cette population, désignée *U*, se compose de *N* unités. Comme le montre la figure 1, nous associons à chaque unité un vecteur correspondant aux résultats de variables cibles potentielles. On observe une partie de ces variables cibles dans les enregistrements administratifs et le reste, dans les enquêtes par sondage. Or, si $\mathbf{D}$ se compose uniquement de variables de registre, on peut alors l'estimer par un simple comptage. Autrement, si $\mathbf{D}$ comprend également les variables échantillonnées, nous utilisons un estimateur par régression générale.

Figure 1. Prototype de base de microdonnées



1re étape : construire des sous-ensembles rectangulaires de microdonnées
Selon notre stratégie d'estimation, nous répartissons la base de données représentée à la figure 1 en quatre sous-ensembles. En dérogeant légèrement à notre notation antérieure, nous les désignons par R, $S_1 \cup S_2$, S_1 et S_2. R correspond à l'enregistrement administratif et contient les variables *R*; $S_1 \cup S_2$ correspond à l'union des premier et deuxième échantillons et contient les variables *R* et *U*; S_1 correspond au premier échantillon et contient les variables *R*, *U* et *V*; S_2 correspond au deuxième échantillon et contient les variables *R*, *U* et *Z*. Or, pour estimer un tableau donné à *m* dimensions, il faut notamment déterminer le sous-ensemble de microdonnées pertinent. Par exemple, si l'estimation concerne uniquement les variables *R*, le sous-ensemble de microdonnées pertinent est R, alors qu'un recoupement entre les variables *V* et *R* doit être estimé à partir de S_1.

2e étape : attribuer des poids de régression
L'étape suivante consiste à attribuer, selon un système de pondération, un ensemble (fixe) de poids à chaque sous-ensemble de microdonnées pour rajuster (globalement) en fonction de l'erreur d'échantillonnage et de la non-réponse et pour répondre à certaines exigences en matière de cohérence. À cette fin, nous devons d'abord calculer les poids de départ pour chaque sous-ensemble de microdonnées.

Pour R, S_1 et S_2, on calcule aisément les poids de départ. Pour R, ils sont égaux à 1, car ce sous-ensemble correspond à un enregistrement complet; pour S_1 et S_2, ils sont égaux à l'inverse des probabilités d'inclusion de premier ordre (nettes) respectives du premier et du deuxième échantillon. Pour $S_1 \cup S_2$, on peut calculer les poids de départ de diverses façons, dont les suivantes. Supposons que π_{1i} et π_{2i} désignent les probabilités d'inclusion de premier ordre de la i^e unité, $i \in U$, à l'égard respectivement de S_1 et S_2, et définissent $d_i^* = \lambda \pi_{1i}^{-1}$ pour $i \in S_1 \setminus S_2$, $d_i^* = (1-\lambda)\pi_{2i}^{-1}$ pour $i \in S_2 \setminus S_1$ et $d_i^* = \lambda \pi_{1i}^{-1} + (1-\lambda)\pi_{2i}^{-1}$ pour $i \in S_1 \cap S_2$, où $\lambda \in [0,1]$. Le choix de λ peut refléter la confiance qu'on porte à un échantillon plutôt qu'à l'autre. Par exemple, il peut dépendre d'indicateurs correspondant à plusieurs types d'erreur, dont les erreurs d'échantillonnage ou les erreurs dues à la non-réponse. Pour simplifier le choix, on peut prendre λ proportionnel à la taille relative de l'échantillon de S_1 par rapport à S_2.

Après avoir calculé les poids de départ, il faut élaborer une stratégie pour attribuer des poids de régression (préliminaires) à chacun des sous-ensembles de microdonnées. Une stratégie évidente consiste à pondérer d'abord $S_1 \cup S_2$ en utilisant les variables *R* pertinentes dans le système de pondération et à pondérer ensuite S_1 et S_2. Pour S_1 et S_2, on peut utiliser les variables *R* et *U* ainsi que des recoupements entre les variables *R* et *U* dans les systèmes de pondération.

3e étape : établir une estimation cohérente d'un ensemble de distributions
Un ensemble de poids de départ et un ensemble de poids de régression préliminaires sont liés à chaque sous-ensemble de microdonnées. Les poids de régression sont calculés d'après un système de pondération soigneusement établi pour corriger l'erreur d'échantillonnage et la non-réponse et pour assurer une certaine cohérence. Dans le reste du présent exposé, ces systèmes de pondération sont appelés systèmes de pondération globaux afin de les distinguer d'autres systèmes de pondération individualisés, dont nous parlerons plus loin.

Pour être plus précis, supposons que $\{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_T\}$ est un ensemble de distributions cibles à estimer à partir d'une certaine base de microdonnées. Au moyen des systèmes de pondération globaux, on peut produire des estimations préliminaires pour cet ensemble. Ces estimations sont asymptotiquement non biaisées selon le plan et il existe des formules d'approximation de leurs variances. De plus, bon nombre de ces distributions estimées sont cohérentes entre elles pour deux raisons. Premièrement, toutes les distributions estimées qui sont calculées à partir du même sous-ensemble de microdonnées de base sont automatiquement cohérentes. Deuxièmement, en raison de la cohérence des poids de régression, les distributions estimées qui sont calculées à partir de différents sous-ensembles de microdonnées de base peuvent être cohérentes. Toutefois, comme il n'est souvent possible d'intégrer qu'un nombre limité de variables auxiliaires dans un système de pondération, on ne peut garantir une cohérence absolue entre les estimations calculées à partir de sous-ensembles différents.

Si l'on est disposé à abandonner la pratique répandue qui consiste à utiliser un seul ensemble de poids (de régression) par sous-ensemble de microdonnées, on obtient alors une plus grande cohérence. Puis, étant donné une certaine distribution à estimer, il est nécessaire d'en déterminer toutes les « marges » qui ont déjà été estimées et d'utiliser ces marges comme renseignements auxiliaires dans un processus de pondération répétée. Le système de pondération répétée dont on a besoin, au minimum, pour couvrir ces marges est appelé système minimal de pondération répétée. Supposons que $\{\mathbf{D}_{k_1}, \mathbf{D}_{k_2}, ..., \mathbf{D}_{k_S}\}$ désigne le sous-ensemble de $\{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_T\}$ devant faire l'objet d'une pondération répétée parce qu'on ne peut en établir une estimation cohérente au moyen des systèmes de pondération globaux et supposons en outre que, par exemple, $\mathbf{D}_{k_s}$ du sous-ensemble correspond à $\mathbf{D}_t$ de l'ensemble initial. Le système minimal de pondération répétée de $\mathbf{D}_{k_s}$ se compose alors des termes de pondération $\mathbf{D}_{(1,t)}, \mathbf{D}_{(2,t)}, \ldots, \mathbf{D}_{(t-1,t)}$, où $\mathbf{D}_{(i,t)}$ représente la distribution de fréquences commune entre $\mathbf{D}_i$ et $\mathbf{D}_t$, c'est-à-dire que $\mathbf{D}_{(i,t)}$ correspond à la partition $\min(\mathbf{r}_i, \mathbf{r}_t)$.

Avant tout, on doit calculer séparément le système minimal de pondération répétée pour chaque distribution. Étant donné un ordre spécifique de l'ensemble de distributions cibles, ces systèmes de pondération répétée sont définis de manière unique. Toutefois, un ordre différent suppose généralement un ensemble différent de systèmes de pondération répétée. Dans la prochaine section, nous illustrerons à l'aide d'un exemple comment on peut résoudre (en grande partie) ce problème d'ordre.

4. UN EXEMPLE NUMÉRIQUE : SAMPLONE

La façon la plus simple d'illustrer notre stratégie d'estimation consiste à présenter un exemple numérique. Supposons que la taille de la population est N = 1 000 personnes. Un enregistrement intégral fournit les résultats des variables région (il y a sept municipalités : Wheaton, Greenham, Newbay, Oakdale, Smokeley, Crowdon, et Mudwater), âge (trois catégories d'âge : jeunes, âge moyen, aînés) et sexe (hommes, femmes). En outre, un échantillon aléatoire simple de taille n = 100 fournit les résultats de la variable emploi (oui, non). En ce qui concerne la région, nous définissons une variable de deuxième ordre en combinant les municipalités de deux provinces, soit Agria (qui comprend Wheaton, Greenham et Newbay) et Induston (qui comprend Oakdale, Smokeley, Crowdon et Mudwater).

Tableau 4.1. région$^{(2)}$ × sexe$^{(1)}$

	Wheaton	Greenham	Neybay	Oakdale	Crowdon	Smokeley	Mudwater	Total
Hommes	70	44	31	36	128	80	122	511
Femmes	74	50	24	25	116	67	133	489
Total	144	94	55	61	244	147	255	1 000

Tableau 4.2. sexe$^{(1)}$ × région$^{(1)}$ × âge$^{(1)}$

Hommes				Femmes			
	Agria	Induston	Total		Agria	Induston	Total
Jeunes	80	146	226	Jeunes	61	148	209
Âge moyen	47	156	203	Âge moyen	57	135	192
Âînés	18	64	82	Âînées	30	58	88
Total	145	366	511	Total	148	341	489

Conformément au modèle conceptuel, nous distinguons quatre suites hiérarchiques de variables de classement (simples) : [région$^{(2)}$, région$^{(1)}$, région$^{(0)}$], [sexe$^{(1)}$, sexe$^{(0)}$], [âge$^{(1)}$, âge$^{(0)}$] et [emploi$^{(1)}$, emploi$^{(0)}$]. À l'aide de ces quatre suites hiérarchiques, on peut construire au maximum 3×2×2×2 = 24 variables de classement (multiples). Nous remarquons que la variable de classement la plus détaillée se compose de 7×3×2×2 = 84 classes et correspond à une partition désignée par $\mathbf{r} = (2 \quad 1 \quad 1 \quad 1)^t$. Notre but consiste à établir une estimation cohérente de l'ensemble de distributions suivant :

- région$^{(2)}$×sexe$^{(1)}$,
- région$^{(1)}$×sexe$^{(1)}$×âge$^{(1)}$,
- région$^{(2)}$×emploi$^{(1)}$,
- emploi$^{(1)}$×âge$^{(1)}$,
- sexe$^{(1)}$×emploi$^{(1)}$×région$^{(1)}$

On peut obtenir les deux premières distributions cibles par de simples comptages de registre. Les résultats sont présentés dans les tableaux 4.1 et 4.2. Les comptages pour la région$^{(2)}$ (tableau 4.1) concordent avec le comptage pour la région$^{(1)}$ (tableau 4.2), ce qui est une condition pour respecter l'exigence en matière de cohérence.

Afin d'estimer le reste des distributions cibles, nous employons la stratégie d'estimation suivante : on adopte d'abord le système de pondération (global) « région$^{(1)}$ + sexe$^{(1)}$ » pour corriger l'erreur d'échantillonnage (et le biais dû à la non-réponse). En faisant abstraction de l'exigence en matière de cohérence, on peut utiliser les poids de régression préliminaires ainsi calculés pour obtenir les estimations (préliminaires) désirées. Les résultats figurent à l'annexe I. Si ces tableaux sont cohérents entre eux, une comparaison entre ces tableaux d'une part et les comptages de registre d'autre part (tableaux 4.1 et 4.2) révèle néanmoins plusieurs incohérences à l'égard des variables âge$^{(1)}$ et région$^{(2)}$, par exemple.

Pour pallier ces incohérences, nous appliquons la théorie de la pondération répétée minimale énoncée dans la section précédente. Les résultats sont résumés dans le tableau 4.3. Nous ne traiterons en détail que de la distribution $\mathbf{D}_3$ = région$^{(2)}$×emploi$^{(1)}$. Les mêmes considérations s'appliquent aux autres distributions cibles. Afin de résoudre (en grande partie) le problème de l'ordre, on estime d'abord ses marges unidimensionnelles région$^{(2)}$ et emploi$^{(1)}$. Comme on peut estimer ces marges au moyen du modèle de pondération global sans déroger aux exigences en matière de cohérence, il n'est pas nécessaire de répéter la pondération. Le système minimal de pondération répétée pour $\mathbf{D}_3$ est donc emploi$^{(1)}$ + région$^{(2)}$.

Tableau 4.3. Systèmes de pondération répétée[3]

Distribution de fréquences	Sous-ensemble pertinent	Pondération répétée nécessaire	Systèmes de pondération répétée
région$^{(2)}$×sexe$^{(1)}$	R	Non	-
région$^{(1)}$×sexe$^{(1)}$×âge$^{(1)}$	R	Non	-
région$^{(2)}$	R	Non	-
emploi$^{(1)}$	S	Non	-
région$^{(2)}$×emploi$^{(1)}$	S	Oui	emploi$^{(1)}$ + région$^{(2)}$
emploi$^{(1)}$	S	Non	-
âge$^{(1)}$	R	Non	-
emploi$^{(1)}$×âge$^{(1)}$	S	Oui	emploi$^{(1)}$ + âge$^{(1)}$
sexe$^{(1)}$	R	Non	-
emploi$^{(1)}$	S	Non	-
région$^{(1)}$	R	Non	-
sexe$^{(1)}$×emploi$^{(1)}$	S	Non	-
sexe$^{(1)}$×région$^{(1)}$	R	Non	-
emploi$^{(1)}$×région$^{(1)}$	S	Non	-
sexe$^{(1)}$×emploi$^{(1)}$×région$^{(1)}$	S	Oui	sexe$^{(1)}$×emploi$^{(1)}$ + sexe$^{(1)}$×région$^{(1)}$ + emploi$^{(1)}$×région$^{(1)}$

Les estimations ainsi calculées figurent à l'annexe II. Naturellement, ces estimations sont cohérentes entre elles et à l'égard de tous les comptages de registre.

5. RÉSUMÉ ET SUJETS DE RECHERCHE

Dans la présente communication, nous avons proposé une stratégie d'estimation à partir de sources de données combinées. Nous avons décidé de combiner des sources de données et d'élaborer une nouvelle stratégie d'estimation en raison des pressions gouvernementales en vue de réduire le fardeau de réponse et de répondre aux exigences des utilisateurs en produisant des produits cohérents et interdépendants. Notre stratégie d'estimation est toujours fondée sur des techniques de régression – ou, plus généralement, de calage – mais pas nécessairement sur un seul système de pondération par enquête. Elle comporte trois étapes : 1) construire des ensembles de microdonnées rectangulaires à partir des sources de données combinées, 2) attribuer à chaque ensemble de microdonnées un ensemble (fixe) de poids de régression (ou de calage) selon un système de pondération, et 3) pour chaque tableau cible, rajuster au minimum le système de pondération initial pour obtenir un système de pondération répétée, dit « minimal », adapté à l'exigence en matière de cohérence. Nous avons illustré l'aspect pratique de cette stratégie d'estimation à l'aide d'un exemple fictif.

La stratégie d'estimation proposée est préliminaire et doit faire l'objet d'une étude plus poussée. Nous allons mentionner brièvement certaines difficultés. L'idée d'une pondération répétée pour obtenir des estimations numériques cohérentes suppose l'existence de bases de microdonnées « parfaites », c'est-à-dire composées d'un nombre (modéré) d'ensembles de microdonnées rectangulaires qui sont cohérents entre eux au niveau des microdonnées. Toutefois, la construction de telles bases de microdonnées constitue une tâche très complexe, qui suppose bien des choix difficiles. Nous avons mis au point un (prototype de) logiciel pour mener à bien les trois étapes du processus de pondération répétée. Bien qu'il soit possible de calculer des formules d'approximation de la variance, les expressions obtenues sont plutôt compliquées; nous ne les avons donc pas encore mises en œuvre. Enfin, nous mentionnons deux complications théoriques liées aux variables quantitatives, soit le double rôle des variables quantitatives et le problème des sous-variables. À l'égard de la première, on peut utiliser une variable quantitative, telle que l'âge, à la fois comme variable de classement et comme variable de quantification. Surtout lorsqu'une variable de quantification suppose un nombre fini de valeurs, où chaque valeur correspond à une classe de variable de classement correspondante, le problème de la cohérence devient manifeste. Les exemples du second problème sont souvent formulés en fonction de règles de contrôle, comme

[3] Afin de pallier certaines lacunes techniques, ces systèmes de pondération répétée peuvent différer légèrement de ceux que proposent Renssen et coll. (2001).

« coûts matériels » + « coûts personnels » = « coûts totaux ». On peut alors considérer les variables « coûts matériels » et « coûts personnels » comme des sous-variables des « coûts totaux ». Nous avons déjà mis au point une méthodologie pour pallier ces complications, mais notre recherche se poursuit.

BIBLIOGRAPHIE

Al, P.G. et F.M. Bakker (2000), « Re-engineering Social Statistics by Micro-integration of Different Sources: an Introduction », *Netherlands Official Statistics*, n° 15, p. 4 à 6, numéro spécial, *Integrating Administrative Registers and Household Surveys*.

Bethlehem, J. et W. Keller (1987), « Linear Weighting of Sample Survey Data », *Journal of Official Statistics*, n° 3, p. 141 à 153.

Deville, J.C. et C.E. Särndal (1992), « Calibration Estimators in Survey Sampling », *Journal of the American Statistical Association*, n° 87, p. 376 à 382.

Kroese, A.H., R.H. Renssen et M. Trijssenaar (2000), « Weighting or Imputation : constructing a consistent set of estimates based on data from different sources », *Netherlands Official Statistics*, n° 15, p. 23 à 31, numéro spécial, *Integrating Administrative Registers and Household Surveys*.

Laan van der, P. (2000), « Integrating Administrative Registers and Household Surveys », *Netherlands Official Statistics*, n° 15, p. 7 à 15, numéro spécial, *Integrating Administrative Registers and Household Surveys*.

Renssen, R.H., A.H. Kroese et A.J. Willeboordse (2001), « Aligning Estimates by Repeated Weighting », rapport inédit (BPA H 491-01-TMO), Heerlen (Pays-Bas), Statistics Netherlands.

Renssen, R.H. et N.J. Nieuwenbroek (1997), « Aligning Estimates for Common Variables in two or more Sample Surveys », *Journal of the American Statistical Association*, n° 90, p. 368 à 374.

Willeboordse, A. (2000), « Towards a new Statistics Netherlands. Blueprint for a process oriented organisation structure », rapport inédit, Voorburg (Pays-Bas), Statistics Netherlands.

Willeboordse, A. et W. Ypma (1996), « From Rules to Tools. New Opportunities to Establish Coherence among Statistics », *Proceedings of the Conference on output Databases*, Voorburg, novembre 1996, Statistics Netherlands, Voorburg (Pays-Bas).

Willeboordse, A. et W. Ypma (1998), « Meta Tools in Support of a Corporate Dissemination Strategy », rapport inédit (document de recherche n° 9839), Voorburg (Pays-Bas), Statistics Netherlands.

Annexe I :

Estimations de trois distributions cibles selon le modèle de pondération « région[(1)] + sexe[(1)] ».

Tableau 1. région[(2)] × emploi[(1)]

	Wheaton	Greenham	Neybay	Oakdale	Crowdon	Smokeley	Mudwater	Total
Avec emploi	66	25	32	33	65	77	66	363
Sans emploi	66	73	32	22	123	111	211	637
Total	131	97	65	55	188	187	276	1 000

Tableau 2. emploi[(1)] × âge[(1)]

	Jeunes	Âge moyen	Aînés	Total
Avec emploi	86	255	22	363
Sans emploi	357	106	174	637
Total	443	361	196	1 000

Tableau 3. sexe[(1)] × emploi[(1)] × région[(1)]

	Hommes				Femmes		
	Agria	Induston	Total		Agria	Induston	Total
Avec emploi	48	207	255	Avec emploi	75	34	108
Sans emploi	104	152	256	Sans emploi	67	314	381
Total	152	359	511	Total	141	348	489

Annexe II :

Estimations de trois tableaux cibles selon le système de pondération répétée (voir le tableau 4.3)

Tableau 1. région[(2)] × emploi[(1)]

	Wheaton	Greenham	Neybay	Oakdale	Crowdon	Smokeley	Mudwater	Total
Avec emploi	72	24	27	36	84	60	60	363
Sans emploi	72	70	28	25	160	87	195	637
Total	144	94	55	61	244	147	255	1 000

Tableau 2. emploi[(1)] × âge[(1)]

	Jeunes	Âge moyen	Aînés	Total
Avec emploi	76	270	17	363
Sans emploi	359	125	153	637
Total	435	395	170	1 000

Tableau 3. sexe[(1)] × emploi[(1)] × région[(1)]

	Hommes				Femmes		
	Agria	Induston	Total		Agria	Induston	Total
Avec emploi	46	209	255	Avec emploi	77	31	108
Sans emploi	99	157	256	Sans emploi	71	310	381
Total	145	366	511	Total	148	341	489