

Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques

J.P. Reiter¹

Résumé

L'une des méthodes permettant d'éviter les divulgations consiste à diffuser des ensembles de microdonnées à grande diffusion partiellement synthétiques. Ces ensembles comprennent les unités enquêtées au départ, mais certaines valeurs recueillies, comme celles de nature délicate présentant un haut risque de divulgation ou celles d'identificateurs clés, sont remplacées par des imputations multiples. Bien qu'on recoure à l'heure actuelle à des approches partiellement synthétiques pour protéger les données à grande diffusion, on ne les a pas encore assorties de méthodes d'inférence valides. Le présent article décrit de telles méthodes. Elles sont fondées sur les concepts de l'imputation multiple en vue de remplacer des données manquantes, mais s'appuient sur des règles différentes pour combiner les estimations ponctuelles et les estimations de la variance. Ces règles de combinaison diffèrent aussi de celles élaborées par Raghunathan, Reiter et Rubin (2003) pour les ensembles de données entièrement synthétiques. La validité de ces nouvelles règles est illustrée au moyen d'études par simulation.

Mots clés : Confidentialité; divulgation; imputation multiple; données synthétiques.

1. Introduction

Lors de la diffusion de données au grand public, les organismes statistiques s'efforcent de fournir des données détaillées sans divulguer les renseignements de nature délicate fournis par les répondants. Pour réduire le risque de divulgation, ils modifient habituellement les données originales avant leur diffusion, par exemple, en recodant les variables, en permutant certaines données ou en ajoutant un bruit aléatoire aux valeurs des variables (Willenborg et de Waal 2001). Cependant, ces méthodes peuvent fausser les liens entre les variables incluses dans l'ensemble de données. Elles compliquent aussi l'analyse car, pour analyser correctement des données perturbées, les utilisateurs devraient suivre les méthodes fondées sur la vraisemblance décrites par Little (1993) ou les modèles d'erreur de mesure décrits par Fuller (1993). Or, ces techniques sont difficiles à appliquer à l'estimation de paramètres non standards et obligent parfois les analystes à maîtriser de nouvelles méthodes statistiques et de nouveaux logiciels spécialisés.

Rubin (1993) a proposé une autre approche qui consiste à : diffuser des ensembles de données entièrement synthétiques complètement constitués de valeurs produites par imputation multiple au lieu de valeurs réelles. Cette méthode permet de protéger les renseignements confidentiels, puisque l'identification des unités et de leurs données de nature délicate devient difficile quand les données diffusées ne sont pas les valeurs réelles recueillies. En outre, s'ils appliquent les méthodes d'imputation et d'estimation appropriées fondées sur les concepts de l'imputation multiple (Rubin 1987), les utilisateurs des données peuvent obtenir des inférences valides grâce à des

méthodes et des logiciels statistiques standards, applicables aux données complètes. Ces inférences peuvent être obtenues par les méthodes de Raghunathan et coll. (2003), qui reposent sur des règles de combinaison des estimations ponctuelles et des estimations de la variance différentes de celles formulées par Rubin (1987). D'autres discussions et variantes des méthodes fondées sur des données synthétiques figurent dans Little (1993), Fienberg, Steele et Makov (1996), Fienberg, Makov et Steele (1998), Dandekar, Cohen et Kirkendall (2002a), Dandekar, Domingo-Ferrer et Sebe (2002b), Franconi et Stander (2002, 2003), Polettini, Franconi et Stander (2002), Polettini (2003) et Reiter (2002, 2003).

Aucun fournisseur de données n'a encore utilisé l'approche entièrement synthétique au stade de la production, mais certains ont adopté une variante, à savoir la diffusion d'ensembles de données partiellement synthétiques comprenant un mélange de valeurs réelles et de valeurs obtenues par imputation multiple. Par exemple, pour protéger les données de la U.S. Survey of Consumer Finances, le U.S. Federal Reserve Board remplace les valeurs monétaires présentant un risque élevé de divulgation par des imputations multiples, puis diffuse un mélange de ces valeurs imputées et des valeurs recueillies non remplacées (Kennickel 1997). Abowd et Woodcock (2001) ont adopté une autre approche partiellement synthétique pour protéger des ensembles de données longitudinales couplées. Ils remplacent toutes les valeurs de certaines variables de nature délicate par des imputations multiples, mais laissent les valeurs d'autres variables telles quelles. Une troisième approche est appliquée par Liu et Little (2002), qui développent un algorithme pour simuler plusieurs valeurs d'identificateurs clés pour certaines

1. J.P. Reiter, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708-0251. Courriel : jerry@stat.duke.edu.

unités. Toutes ces approches partiellement synthétiques sont intéressantes, car elles promettent d'offrir nombre d'avantages des approches entièrement synthétiques, comme assurer la confidentialité des données en permettant aux utilisateurs de faire des inférences sans être obligés de maîtriser des méthodes ou des logiciels statistiques compliqués, tout en réduisant la sensibilité à la spécification des modèles d'imputation.

Bien que des ensembles de données partiellement synthétiques fassent l'objet de grande diffusion, la littérature n'offre pas de renseignements techniques sur la façon de produire des inférences à partir de ces ensembles. À première vue, il peut sembler correct d'utiliser les méthodes d'inférence proposées par Rubin (1987) en cas d'imputation multiple pour remplacer des données manquantes. Malheureusement, comme nous le montrons dans le présent article, ces méthodes produisent parfois des estimations biaisées de la variance. Qui plus est, comme nous le montrons aussi, les méthodes élaborées par Raghunathan et coll. (2003) pour l'analyse des données entièrement synthétiques ne sont pas valides lorsqu'appliquées à des données partiellement synthétiques. De nouvelles méthodes d'inférence sont donc nécessaires.

Le présent article décrit les méthodes d'inférence à partir d'ensembles de données partiellement synthétiques obtenus par imputation multiple. L'élaboration de ces méthodes fournit aussi des instructions pour la production de données partiellement synthétiques. La présentation de l'article est la suivante. La section 2 décrit les nouvelles méthodes d'inférence. La section 3 montre l'élaboration de ces méthodes dans un cadre bayésien et décrit les conditions dans lesquelles les inférences résultantes devraient être valides du point de vue fréquentiste. La section 4 décrit les études en simulation qui illustrent la validité de ces méthodes, ainsi que l'inefficacité des règles concurrentes de combinaison de plusieurs estimations ponctuelles et estimations de la variance. La section 5 contient les conclusions et des suggestions quant aux futurs travaux de recherche.

2. Inférence à partir d'ensembles de données partiellement synthétiques obtenues par imputation multiple

Soit $I_j = 1$ si l'on sélectionne l'unité j dans l'enquête originale et $I_j = 0$ autrement. Soit $I = (I_1, \dots, I_N)$. Soit Y_{obs} la matrice de dimensions $n \times p$ de données d'enquête recueillies (réelles) pour les unités pour lesquelles $I_j = 1$; soit Y_{nobs} la matrice de dimensions $(N-n) \times p$ de données d'enquête non observées pour les unités pour lesquelles $I_j = 0$; et soit $Y = (Y_{\text{obs}}, Y_{\text{nobs}})$. Par souci de simplicité, nous supposons que toutes les unités échantillonnées répondent complètement à l'enquête. Soit X la matrice de dimensions $N \times d$ de variables de plan de sondage pour l'ensemble des N unités de la population, comme des indicateurs de strate ou de grappe ou des

mesures de taille. Nous supposons qu'on connaît approximativement ces renseignements sur le plan de sondage pour toutes les unités de la population. Ils peuvent provenir, par exemple, des dossiers du recensement ou de la base de sondage.

L'organisme qui diffuse des données synthétiques, appelé dans la suite de l'article l'imputeur, construit des ensembles de données synthétiques d'après les données observées, $D = (X, Y_{\text{obs}}, I)$, selon un processus en deux volets. En premier lieu, il sélectionne parmi les données observées les valeurs qui seront remplacées par des données imputées. En deuxième lieu, il impute les nouvelles valeurs pour remplacer les valeurs sélectionnées. Soit $Z_j = 1$ si l'on sélectionne l'unité j pour le remplacement de certaines de ses valeurs observées par des valeurs synthétiques et soit $Z_j = 0$ pour les unités pour lesquelles aucune donnée n'est modifiée. Soit $Z = (Z_1, \dots, Z_n)$. Soit $Y_{\text{rep},i}$ toutes les valeurs imputées (remplacées) dans le i^{e} ensemble de données synthétiques et soit Y_{nrep} l'ensemble des valeurs non modifiées (non remplacées) de Y_{obs} . Nous supposons que les $Y_{\text{rep},i}$ sont générées à partir de la loi prédictive a posteriori bayésienne de $(Y_{\text{rep},i} | D, Z)$. Les valeurs comprises dans Y_{nrep} sont les mêmes pour tous les ensembles de données synthétiques. Alors, chaque ensemble de données synthétique, d_i , comprend $(X, Y_{\text{rep},i}, Y_{\text{nrep}}, I, Z)$. Les imputations sont réalisées indépendamment $i = 1, \dots, m$ fois pour produire m ensembles distincts de données synthétiques. Enfin, ces ensembles sont diffusés au public.

Les valeurs contenues dans Z peuvent, et il en est fréquemment ainsi, dépendre des valeurs contenues dans D . Par exemple, l'imputeur peut choisir de ne simuler des variables ou des identificateurs de nature délicate que pour les unités de l'échantillon présentant une combinaison rare d'identificateurs; ou bien, il peut ne remplacer que les valeurs de revenu supérieures à 100 000 \$ par des valeurs imputées. Pour éviter d'introduire un biais, l'imputeur devrait tenir compte de ce genre de sélection en procédant à l'imputation à partir de la loi prédictive a posteriori de Y pour les unités pour lesquelles $Z_j = 1$. En pratique, il peut le faire en utilisant uniquement les unités pour lesquelles $Z_j = 1$ comme source de données pour rechercher les lois a posteriori pour l'imputation. L'utilisation de toutes les unités pour lesquelles $I_j = 1$ peut produire des estimations biaisées ou des intervalles de confiance plus larges avec des taux de couverture exagérément prudents, comme l'illustrent les simulations présentées à la section 4.

À partir de ces ensembles de données synthétiques, l'utilisateur des données diffusées au grand public, appelé dans la suite de l'article l'analyste, veut faire des inférences au sujet d'un paramètre à estimer $Q = Q(X, Y)$, où la notation $Q(X, Y)$ signifie que Q est une fonction de (X, Y) . Par exemple, Q pourrait représenter la moyenne de population de Y ou les coefficients de régression de population de Y sur X . Pour chaque ensemble de données synthétiques d_i , l'analyste estime Q au moyen d'un

estimateur ponctuel q et la variance de q au moyen d'un estimateur v . Nous supposons que l'analyste détermine les estimateurs q et v comme si les données synthétiques étaient en fait des données recueillies à partir d'un échantillon aléatoire de (X, Y) d'après le plan de sondage réel utilisé pour générer I .

Pour $i = 1, \dots, m$, soit q_i et v_i les valeurs respectives de q et v dans l'ensemble de données synthétiques d_i . Sous certaines conditions, que nous décrivons à la section 3, l'analyste peut obtenir des inférences valides pour la grandeur scalaire Q en combinant les q_i et v_i . Précisément, les quantités qui suivent sont nécessaires pour les inférences :

$$\bar{q}_m = \sum_{i=1}^m q_i / m \quad (1)$$

$$b_m = \sum_{i=1}^m (q_i - \bar{q}_m)^2 / (m-1) \quad (2)$$

$$\bar{v}_m = \sum_{i=1}^m v_i / m. \quad (3)$$

L'analyste peut alors utiliser \bar{q}_m pour estimer Q et

$$T_p = b_m / m + \bar{v}_m \quad (4)$$

pour estimer la variance de \bar{q}_m . Si q est une fonction de (X, Y_{repl}, I) uniquement et n'est une fonction d'aucune valeur imputée, les inférences à partir des données synthétiques sont identiques aux inférences à partir des données observées; autrement dit, $q_i = q_{\text{obs}}$ et $v_i = v_{\text{obs}}$ pour tout i , et $b_m = 0$. Quand n est grand, les inférences concernant la grandeur scalaire Q peuvent se fonder sur des lois de Student à $v_p = (m-1)(1+r_m^{-1})^2$ degrés de liberté, où $r_m = (m^{-1}b_m/\bar{v}_m)$. Dans de nombreux cas, r_m^{-1} et, donc, v_p est suffisamment grand pour qu'une loi normale représente une approximation adéquate de la loi de Student. Nous ne présentons pas ici les extensions au cas où Q est multivarié.

T_p diffère de l'estimateur de la variance pour l'imputation multiple de données manquantes, $T_m = (1+1/m)b_m + \bar{v}_m$ (Rubin 1987). Dans le contexte de données partiellement synthétiques, \bar{v}_m estime $\text{Var}(q_{\text{obs}})$ et b_m/m estime la variance supplémentaire due à l'utilisation d'un nombre fini d'imputations. Dans le contexte de données manquantes, \bar{v}_m et b_m/m ont la même interprétation, mais un b_m supplémentaire est nécessaire pour obtenir la moyenne sur le mécanisme de non-réponse (Rubin 1987, chapitre 4). Cette moyenne supplémentaire n'est pas nécessaire dans le cas de données partiellement synthétiques, puisque le mécanisme de sélection Z , qui est déterminé par l'imputeur, n'est pas traité comme étant stochastique.

T_p diffère aussi de l'estimateur de la variance pour l'analyse de données entièrement synthétiques, $T_s = (1+1/m)b_m - \bar{v}_m$ (Raghunathan et coll. 2003). Pour générer des données entièrement synthétiques, on sélectionne

de nouvelles unités à partir de la ou des bases de sondage pour chaque ensemble de données synthétiques, puis on impute leurs valeurs. Comme l'ont montré Raghunathan et coll. (2003), ce processus de rééchantillonnage et d'imputation fait que $b_m - \bar{v}_m$ est une estimation appropriée de $\text{Var}(q_{\text{obs}})$. Pour les données partiellement synthétiques, on utilise les unités originales pour chaque ensemble de données, si bien que \bar{v}_m est une estimation appropriée de $\text{Var}(q_{\text{obs}})$.

3. Justification des nouvelles règles de combinaison

La présente section illustre un calcul bayésien des inférences décrites à la section 2 et les conditions sous lesquelles ces inférences sont valides du point de vue fréquentiste. Ces résultats sont fondés sur la théorie élaborée par Raghunathan et coll. (2003) et la suite de près.

3.1 Calcul bayésien

Pour ce calcul, nous supposons que l'analyste et l'imputeur utilisent le même modèle bayésien. Nous pouvons décomposer la loi a posteriori de $(Q|d^m)$, où $d^m = \{d_1, d_2, \dots, d_m\}$, de la façon suivante

$$f(Q|d^m) = \int f(Q|d^m, D, B) f(D|d^m, B) f(B|d^m) dD dB \quad (5)$$

où $B = \text{Var}(q_i|D, Z)$. En ce qui concerne $f(D|d^m, B)dD$, l'intégration se fait uniquement sur les valeurs de Y_{obs} qui sont remplacées par des valeurs imputées; les composantes (X, Y_{repl}, I) de D restent fixes. Sachant D , les données synthétiques sont sans pertinence, de sorte que $f(Q|d^m, D, B) = f(Q|D)$. Nous supposons que les hypothèses asymptotiques bayésiennes classiques sont vérifiées, si bien que $f(Q|D) \sim N(q_{\text{obs}}, v_{\text{obs}})$, où q_{obs} et v_{obs} sont les moyenne et variance a posteriori de Q déterminées en utilisant D .

L'intégration de (5) sur D donne $f(Q|d^m, B)$. Puisque seules q_{obs} et v_{obs} sont nécessaires pour les inférences au sujet de $(Q|D)$, pour $f(D|d^m, B)$, il est suffisant de déterminer $f(q_{\text{obs}}, v_{\text{obs}}|d^m, B)$. Nous supposons que les imputations sont faites de telle façon que, pour tout i , $(q_i|D, B) \sim N(q_{\text{obs}}, B)$ et $(v_i|D, B) \sim (v_{\text{obs}}, \ll B)$. Ici, la notation $F \sim (G, \ll H)$ signifie que la variable aléatoire F suit une loi dont l'espérance de G et sa variabilité sont nettement plus faibles que pour H . En réalité, v_i est habituellement centrée à une valeur supérieure à v_{obs} , puisque les données synthétiques introduisent une incertitude due au tirage des valeurs des paramètres. Pour les échantillons dont la taille n est grande, ce biais devrait être minimal. L'hypothèse selon laquelle $E(q_i|D, B) = q_{\text{obs}}$ devrait être raisonnable si les imputations sont tirées de la loi a posteriori correcte de Y pour les unités pour lesquelles $Z_j = 1$.

Si l'on suppose que les lois a priori de q_{obs} et v_{obs} , sont uniformes, la théorie bayésienne classique implique que $(q_{\text{obs}} | d^m, B) \sim N(\bar{q}_m, B/m)$ et $(v_{\text{obs}} | d^m, B) \sim (\bar{v}_m, \ll B/m)$. Donc, les moyenne et variance a posteriori de $(Q | d^m, B)$ sont

$$\begin{aligned} E(Q | d^m, B) &= E(E(Q | D, d^m, B) | d^m, B) \\ &= E(q_{\text{obs}} | d^m, B) = \bar{q}_m \end{aligned} \quad (6)$$

$$\begin{aligned} \text{Var}(Q | d^m, B) &= E(\text{Var}(Q | D, d^m, B) | d^m, B) \\ &\quad + \text{Var}(E(Q | D, d^m, B) | d^m, B) \\ &= \bar{v}_m + B/m. \end{aligned} \quad (7)$$

Puisque toutes les convolutions portent sur des lois normales $f(Q | d^m, B) \sim N(\bar{q}_m, \bar{v}_m + B/m)$.

Pour intégrer cette loi sur $f(B | d^m)$, nous utilisons le fait que $((m-1)b_m B^{-1} | d^m) \sim \chi_{m-1}^2$ et, suivant l'approximation de Rubin (1987), nous ajustons les deux premiers moments de $\bar{v}_m + B/m$ sur la moyenne quadratique d'une variable aléatoire. L'approximation résultante de la loi a posteriori de Q est $(Q | d^m) \sim t_{v_p}(\bar{q}_m, T_p)$, où v_p est telle que définie à la section 2.

3.2 Validité de la randomisation

Pour que les inférences fondées sur les équations (1) à (4) aient des propriétés fréquentistes valides, nous devons imposer deux conditions. Premièrement, l'analyste doit utiliser des estimateurs q et v valides du point de vue de la randomisation. Autrement dit, lorsqu'on applique q et v à D pour obtenir q_{obs} et v_{obs} , $(q_{\text{obs}} | X, Y) \sim N(Q, U)$ et $(v_{\text{obs}} | X, Y) \sim (U, \ll U)$, où la loi pertinente est celle de I . Deuxièmement, les méthodes de génération de données synthétiques doivent être correctes au sens de Rubin (1987). Plus précisément, les méthodes de génération des données doivent satisfaire les conditions suivantes :

C1: Si l'on fait la moyenne sur les imputations de $Y_{\text{rep},i}$, il est nécessaire que

- (i) $(q_i | X, Y, I, Z) \sim N(q_{\text{obs}}, B)$;
- (ii) $(b_m | X, Y, I, Z) \sim (B, \ll B)$; et
- (iii) $(\bar{v}_m | X, Y, I, Z) \sim (v_{\text{obs}}, \ll B/m)$, où $B = \text{Var}(q_i | X, Y, I, Z)$.

C2: Si l'on fait la moyenne sur les mécanismes d'échantillonnage et de remplacement $(I, Z | X, Y)$, il est nécessaire que $(B | X, Y) \sim (B_0, \ll U)$ où $B_0 = E(b_m | X, Y)$.

Essentiellement, ces conditions exigent que les données synthétiques soient générées de sorte que les q_i soient sans biais pour q_{obs} , b_m soit sans biais pour B_0 , et \bar{v}_m soit sans biais pour v_{obs} . Une discussion plus approfondie du procédé d'imputation approprié figure dans Rubin (1987).

Partant de ces hypothèses, il suit que

$$\begin{aligned} E(\bar{q}_m | X, Y) &= E(E(\bar{q}_m | X, Y, I, Z) | X, Y) \\ &= E(q_{\text{obs}} | X, Y) = Q \end{aligned} \quad (8)$$

$$\begin{aligned} \text{Var}(\bar{q}_m | X, Y) &= E(\text{Var}(\bar{q}_m | X, Y, I, Z) | X, Y) \\ &\quad + \text{Var}(E(\bar{q}_m | X, Y, I, Z) | X, Y) \\ &= E(B | X, Y) / m + \text{Var}(q_{\text{obs}} | X, Y) = B_0 / m + U. \end{aligned} \quad (9)$$

Puisque nous supposons que $(q_{\text{obs}} | X, Y)$ et $(q_i | X, Y, I, Z)$ suivent une loi normale, il s'ensuit que $(\bar{q}_m | X, Y) \sim N(Q, B_0 / m + U)$.

Si C1 et C2 sont vérifiées, T_p est un estimateur sans biais de $B_0 / m + U$. L'approximation de t est justifiée par la méthode décrite dans Rubin (1987). Plus précisément, l'approximation de t découle du fait que $((m-1)b_m B_0^{-1} | X, Y) \sim \chi_{m-1}^2$, et que le nombre de degrés de liberté d'une variable aléatoire suivant la loi du chi carré est égal au double du carré de son espérance sur sa variance.

4. Études en simulation

La présente section illustre la validité des nouvelles règles de combinaison susmentionnées, ainsi que l'inefficacité de T_m et de T_s en tant qu'estimateurs de la variance grâce à des études en simulation de stratégies partiellement synthétiques. La section 4.1 décrit deux études où l'imputeur génère des données synthétiques uniquement pour certaines unités. La section 4.2 décrit une étude où l'imputeur génère des données synthétiques pour toutes les valeurs d'une variable d'enquête, en laissant les autres variables à leurs valeurs observées. Les simulations sont basées ici sur des données artificielles et des lois a posteriori correctes pour les imputations. Naturellement, dans les conditions réelles, l'imputeur ne connaît habituellement pas le modèle d'imputation correct et doit l'estimer en s'appuyant sur des données observées et l'expertise disponible du domaine spécialisé. Pour toutes les simulations, les tailles de population sont considérées comme étant infinies afin de pouvoir omettre les facteurs de correction pour population finie.

4.1 Imputation pour certaines unités

L'imputeur peut décider de remplacer les valeurs observées pour quelques unités parmi les données recueillies, puis de diffuser un mélange de valeurs imputées et observées. Nous utilisons cette stratégie dans deux simulations simplistes quoiqu'illustratives, la première comportant une variable et la seconde, quatre variables.

4.1.1 Simulations à l'aide d'une seule variable

Chaque ensemble de données observé, D , comprend $n=100$ valeurs tirées aléatoirement à partir de

$Y \sim N(0, 10^2)$. Nous utilisons deux scénarios distincts pour spécifier les unités pour lesquelles $Z_j = 1$, afin de générer deux ensembles de données partiellement synthétiques pour chaque D . Le premier scénario, appelé « aléatoire », consiste à remplacer Y pour 20 unités échantillonnées aléatoirement à partir de D . Le deuxième, appelé « grand Y » consiste à remplacer Y uniquement pour les unités pour lesquelles $Y_j > 10$.

Pour chaque D , et pour chaque scénario, il existe $m = 5$ ensembles de données synthétiques, $d_i = (Y_{\text{rep}, i}, Y_{\text{rep}}, I, Z)$, pour $i = 1, \dots, 5$. Les $Y_{\text{rep}, i}$ sont générés selon une technique bootstrap bayésienne (Rubin 1987, pages 123–124) qui consiste à tirer des valeurs de Y à partir d'un groupe donneur de valeurs sélectionnées de Y_{obs} . Soit Y_{elig} le vecteur de dimensions $n_0 \times 1$ des valeurs de Y_{obs} qui constituent le groupe donneur. Soit $n_{\text{rep}} = \sum_{j=1}^{100} Z_j$. Le procédé bootstrap bayésien se déroule comme suit :

1. Tirer $(n_0 - 1)$ nombres aléatoires uniformes. Les classer par ordre ascendant. Appliquer à ces nombres ordonnés l'étiquetage $a_0 = 0, a_1, a_2, \dots, a_{n_0-1}, a_{n_0} = 1$.
2. Tirer n_{rep} nombres aléatoires uniformes, $u_1, u_2, \dots, u_j, \dots, u_{n_{\text{rep}}}$. Pour chacun de ces u , imputer $Y_{\text{elig}, j}$ quand $a_{j-1} < u \leq a_j$.

Il est peu probable qu'on utilise ce bootstrap bayésien pour imputer des données dans des conditions réelles, puisque les ensembles de données contiennent plus d'une variable. Nous l'employons ici parce qu'il produit des imputations simples, appropriées pour la présente illustration.

Comme nous l'avons mentionné à la section 2, la loi prédictive a posteriori correcte est $f(Y|D, Z)$ et non $f(Y|D)$. Par conséquent, le groupe donneurs, Y_{elig} , devrait être égal à l'ensemble $\{Y_j : Z_j = 1\}$ que nous nommons « SÉLECTION ». Aux fins de comparaison, nous imputons aussi des valeurs synthétiques au moyen de

l'ensemble de donneurs $\{Y_j : I_j = 1\}$ que nous nommons « TOUT ». Les imputations fondées sur TOUTES les unités ne satisfont pas la condition C1 de la section 3.2, puisque $E(q_i | X, Y, I, Z) = (\sum_{j=1}^{100-n_{\text{rep}}} y_{\text{rep}, j} + n_{\text{rep}} \bar{y}_{\text{obs}}) / n \neq \bar{y}_{\text{obs}}$, tandis que celles fondées sur les unités SÉLECTIONNÉES sont correctes.

Le tableau 1 résume les résultats de 5 000 exécutions de cette simulation. Aussi bien pour le scénario « aléatoire » que pour le scénario « grand Y », les moyennes de \bar{q}_5 fondées sur les unités SÉLECTIONNÉES sont approximativement égales à la moyenne de q_{obs} . Dans le cas du scénario aléatoire, \bar{q}_5 fondé sur TOUTES les unités est également sans biais, car $E(\bar{y}_{\text{rep}} | X, Y, I) = q_{\text{obs}}$ lorsque la moyenne est calculée sur Z (qui est en fait stochastique dans ce scénario). Cependant, si l'on utilise TOUTES les unités dans le scénario « grand Y », \bar{q}_5 présente un biais par défaut important, parce que les valeurs imputées ne sont pas contraintes d'être supérieures à 10 lorsqu'on utilise TOUTES les unités.

Aussi bien dans le scénario « aléatoire » que le scénario « grand Y », 94,5 % des 5 000 intervalles de confiance à 95 % synthétiques fondés sur T_p et les unités SÉLECTIONNÉES contiennent la valeur zéro. Ce taux est identique au taux de couverture de 94,5 % des intervalles de confiance fondés sur les données observées ($q_{\text{obs}} \pm 1.96\sqrt{v_{\text{obs}}}$). Les taux nominaux sont inférieurs à 95 % à cause de l'erreur de simulation. Les 2 à 3 % d'écart entre les moyennes de T_p et de $\text{Var}(\bar{q}_5)$ équivalent à peu près à l'écart entre les moyennes de v_{obs} et de $\text{Var}(q_{\text{obs}})$. L'estimateur habituel de la variance sous imputation multiple, T_m , a tendance à surestimer $\text{Var}(\bar{q}_5)$, ce qui produit un taux de couverture des intervalles de confiance exagérément prudent et montre que T_m n'est pas l'estimateur approprié de la variance lorsqu'on analyse des données partiellement synthétiques correctement imputées.

Tableau 1
Résultats des simulations pour l'imputation des valeurs d'une seule variable

Scénario et méthode d'imputation	Moy. \bar{q}_5	Var \bar{q}_5	Moy. T_p	Moy. T_m	Couverture des IC à 95 %	
					En utilisant T_p	En utilisant T_m
$Z_j = 1$ pour 20 unités sélectionnées aléatoirement						
SÉLECTION	0,024	1,097	1,067	1,420	94,5%	96,7%
TOUT	0,020	1,233	1,044	1,281	92,6%	94,9%
$Z_j = 1$ pour les unités pour lesquelles $Y_j > 10$						
SÉLECTION	0,016	1,031	1,011	1,068	94,5%	95,0%
TOUT	-2,383	0,796	0,736	0,921	20,7%	28,8%
Résultats pour les données observées*	0,016	1,021	1,000		94,5%	

* Les titres de colonne ne s'appliquent pas à cette ligne. La moyenne de $q_{\text{obs}} = 0,016$, la variance de $q_{\text{obs}} = 1,021$, la moyenne de $v_{\text{obs}} = 1,000$ et 94,5 % des 5 000 intervalles de confiance à 95 % pour les données observées contiennent la valeur zéro.

Si les imputations se fondent sur TOUTES les unités – une méthode d'imputation incorrecte – dans le cas du scénario « aléatoire », T_p présente un biais par défaut et 92,6 % seulement des intervalles de confiance à 95 % synthétiques contiennent la valeur zéro. L'utilisation de T_m fait passer le taux de couverture à 95 %, ce qui donne à penser qu'il est plus sûr d'utiliser T_m que T_p lorsqu'on utilise TOUTES les unités pour l'imputation. Les intervalles de confiance fondés sur la méthode TOUT et sur T_m sont, en moyenne, plus grands que ceux fondés sur la méthode SÉLECTION et sur T_p . Ce résultat illustre l'avantage qu'il y a à conditionner sur Z pour obtenir des imputations correctes, même si le scénario utilisé pour fixer $Z_j = 1$ ne dépend pas des valeurs de Y .

L'estimateur de la variance pour les données entièrement synthétiques, T_s , qui n'est pas présenté au tableau 1, est négatif pour chacune des 5 000 simulations pour les deux scénarios et pour les deux méthodes d'imputation. Manifestement, bien qu'il soit valide pour les données entièrement synthétiques (Raghuathan et coll. 2003), T_s n'est en général pas approprié pour les données partiellement synthétiques.

4.1.2 Simulations au moyen de quatre variables

Chaque ensemble de données observé, D , comprend $n = 200$ valeurs de quatre variables, (Y_1, Y_2, Y_3, Y_4) , générées comme suit : $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \Sigma)$, où Σ est telle que toutes les variances sont égales à l'unité et que toutes les covariances sont égales à 0,5, et $(y_4 | y_1, y_2, y_3) \sim N(10y_1 + 7y_2 + 4y_3, 25^2)$. Pour fixer les idées, nous pouvons considérer la variable Y_1 comme étant un identificateur clé et Y_4 comme étant une variable de nature délicate. Le plan est de simuler des valeurs de la variable de nature délicate Y_4 pour toutes les unités ayant une valeur « inhabituelle » de l'identificateur clé, définie comme étant $Y_1 > 1$. Donc, Y_{rep} comprend les valeurs échantillonnées de (Y_1, Y_2, Y_3) et les valeurs de Y_4 pour les unités pour lesquelles $Y_1 \leq 1$. Habituellement, environ 30 unités par ensemble de données observé sont telles que $Y_1 > 1$.

Comme précédemment, nous examinons deux scénarios pour déterminer la loi prédictive a posteriori pour les imputations. La méthode SÉLECTION consiste à utiliser uniquement les unités pour lesquelles $Z_j = 1$ comme sources de données pour les lois a posteriori tandis que la méthode TOUT consiste à utiliser toutes les unités observées. Sous chaque scénario, pour procéder aux imputations, i) nous tirons des valeurs des paramètres de la régression de Y_4 sur (Y_1, Y_2, Y_3) à partir de leur loi a posteriori qui est estimée en utilisant les unités SÉLECTIONNÉES ou TOUTES les unités et ii) nous tirons des valeurs de Y_4 pour les unités pour lesquelles $Z_j = 1$ en nous servant des valeurs tirées pour les paramètres. En tout, $m = 5$ ensembles de données synthétiques sont générés pour chaque ensemble de données observé D .

Les paramètres d'intérêt incluent β , le coefficient de régression de Y_1 dans la régression linéaire de Y_4 sur

(Y_1, Y_2, Y_3) ; α , le coefficient de régression de Y_4 dans la régression de Y_1 sur (Y_2, Y_3, Y_4) et \bar{Y}_4 , la moyenne de population de Y_4 . Pour les inférences au sujet de β et α , q est l'estimateur par les moindres carrés ordinaires habituel et v est l'estimateur de la variance. Pour les inférences au sujet de \bar{Y}_4 , q est la moyenne d'échantillon et v est l'erreur-type.

Le tableau 2 résume les résultats pour 5 000 exécutions de cette simulation. Quand les imputations sont fondées sur les unités SÉLECTIONNÉES, les moyennes de \bar{q}_5 et de T_p s'écartent des moyennes de q_{obs} et $\text{Var}(\bar{q}_5)$ d'une valeur égale ou inférieure à l'erreur de simulation. En outre, les taux de couverture des intervalles de confiance à 95 % synthétiques sont semblables à ceux des intervalles de confiance à 95 % pour les données observées. Les valeurs de T_m sont considérablement plus élevées que celles de $\text{Var}(\bar{q}_5)$, ce qui donne des taux de couverture d'environ 97 %. Les valeurs de T_s , qui ne sont pas présentées au tableau 2, sont négatives pour chacune des 5 000 exécutions de la simulation. Dans l'ensemble, ces résultats confirment ceux présentés à la section 4.1.1 : si les valeurs imputées sont tirées d'une loi a posteriori conditionnée sur Z , les estimations ponctuelles et les estimations d'intervalle fondées sur T_p sont plus exactes que celles fondées sur T_m ou sur T_s .

Même si les imputations fondées sur TOUTES les unités ne sont pas correctes, il est instructif d'examiner les propriétés de T_p et de T_m pour ce genre d'imputation. Les imputeurs pourraient fonder leurs imputations sur toutes les unités observées pour des raisons pratiques, par exemple parce que les unités pour lesquelles $Z_j = 1$ ne fournissent pas suffisamment de données pour ajuster les modèles d'imputation. Les résultats reflètent ceux de la section 4.1.1 : T_p sous-estime $\text{Var}(\bar{q}_5)$, ce qui donne des taux de couverture d'environ 94 %, tandis que l'utilisation de T_m porte les taux de couverture à environ 96 %, principalement à cause du biais par excès dans T_m . De nouveau, ces résultats donnent à penser que, si les imputeurs fondent effectivement leurs imputations sur toutes les unités observées bien qu'il n'en existe que quelques-unes pour lesquelles $Z_j = 1$, il est plus prudent que les analystes utilisent T_m au lieu de T_p pour estimer la variance. Comme nous l'avons montré à la section 4.1.1, les intervalles fondés sur TOUTES les unités sont habituellement plus grands que ceux fondés sur les unités SÉLECTIONNÉES, si bien que, dans la mesure du possible, les imputeurs devraient fonder leurs imputations uniquement sur les unités pour lesquelles $Z_j = 1$.

4.2 Imputation de toutes les valeurs de Y pour une variable

Chaque ensemble de données observé comprend $n = 200$ valeurs de quatre variables générées comme suit : $(y_1, y_2, y_3) \sim MVN(\mathbf{0}, \mathbf{I})$, où \mathbf{I} est la matrice d'identité et $(y_4 | y_1, y_2, y_3) \sim N(10y_1 + 10y_2 + 10y_3, 25^2)$. Donc, $Y_{\text{rep}} = (Y_1, Y_2, Y_3)$. Les valeurs de Y_4 sont imputées à partir de la

loi prédictive a posteriori bayésienne de $(Y_4 | Y_{\text{obs}})$, obtenue en ajustant la régression de Y_4 sur (Y_1, Y_2, Y_3) . Toutes les unités sont telles que $Z_j = 1$ et sont utilisées comme sources de données pour les lois a posteriori. Les paramètres à estimer sont les mêmes que ceux décrits à la section 4.1.2.

Le tableau 3 résume les résultats de 5 000 exécutions de la simulation au moyen de $m=5$ ensembles de données partiellement synthétiques. Pour tous les paramètres à estimer, la moyenne de \bar{q}_5 est presque identique à celle de q_{obs} . En outre, les estimations de la variance fondées sur T_p sont proches de la variance réelle de \bar{q}_5 . Le léger biais par excès est dû au fait que \bar{v}_m a tendance à surestimer v_{obs} , comme nous l'avons expliqué à la section 3.1. En moyenne, T_m surestime $\text{Var}(\bar{q}_5)$ d'un facteur supérieur à deux et T_s sousstime fortement $\text{Var}(\bar{q}_5)$ pour α et \bar{Y}_4 . Ces problèmes ne sont pas dus au fait que m est petit, puisqu'ils persistent dans les simulations où m est grand. Même si des erreurs de cette importance ne se produisent pas nécessairement dans d'autres conditions, les résultats de

cet exemple simple indiquent de nouveau que T_m et T_s ne sont, en général, pas appropriés pour analyser des données partiellement synthétiques, surtout quand on produit des valeurs entièrement synthétiques pour certaines variables.

Les imputeurs ont de bonnes raisons de ne diffuser qu'un petit nombre d'ensembles de données synthétiques. Chaque ensemble de données supplémentaire nécessite un espace de stockage supplémentaire et, par dessus tout, la diffusion d'un trop grand nombre d'ensembles de données risque de compromettre le respect de la confidentialité, car des intrus pourraient combiner les valeurs imputées pour obtenir des renseignements sur les valeurs réelles. Le tableau 4 donne les résultats pour des répétitions indépendantes de 5 000 exécutions de la simulation basées sur diverses valeurs de m . Les estimations ponctuelles sont sans biais pour les trois paramètres estimés et ne sont donc pas présentées dans le tableau. Les taux de couverture des intervalles de confiance à 95 % sont proches de 95 % pour toutes les valeurs de m supérieures à deux. La surestimation dans le cas de T_p est de nouveau due à un biais par excès dans \bar{v}_m .

Tableau 2
Résultats des simulations où les valeurs de Y_4 sont imputées pour les unités pour lesquelles $Y_1 > 1$

Type d'inférence	Moy. \bar{q}_5	Var \bar{q}_5	Moy. T_p	Moy. T_m	Couverture des IC à 95 %	
					En utilisant T_p	En utilisant T_m
Paramètre : β						
SÉLECTION	10,02	5,45	5,68	8,97	95,3 %	98,2 %
TOUT	10,04	5,89	5,28	7,57	93,7 %	96,9 %
Données observées*	10,00	4,70			95,5 %	
Paramètre : α						
SÉLECTION	$9,25 \times 10^{-3}$	$4,49 \times 10^{-6}$	$4,76 \times 10^{-6}$	$6,97 \times 10^{-6}$	95,4 %	97,9 %
TOUT	$9,59 \times 10^{-3}$	$5,03 \times 10^{-6}$	$4,75 \times 10^{-6}$	$6,31 \times 10^{-6}$	94,1 %	96,5 %
Données observées*	$9,66 \times 10^{-3}$	$4,26 \times 10^{-6}$			95,4 %	
Paramètre : \bar{Y}_4						
SÉLECTION	$-1,45 \times 10^{-2}$	4,97	5,01	6,09	95,0 %	96,6 %
TOUT	$-1,24 \times 10^{-3}$	5,19	4,82	5,59	93,8 %	95,4 %
Données observées*	$-2,34 \times 10^{-3}$	4,76			94,5 %	

* Les titres de colonne ne s'appliquent pas à cette ligne. Il s'agit de la moyenne de q_{obs} , de la variance de q_{obs} et du pourcentage des intervalles de confiance à 95 % pour les données observées qui couvrent Q .

Tableau 3
Résultats des simulations lors de l'imputation de toutes les valeurs d'une variable

Paramètre à estimer	Moy. q_{obs}	Moy. \bar{q}_5	Var q_{obs}	Var \bar{q}_5	Moy. T_p	Moy. T_m	Moy. T_s
β	9,9500	9,9400	3,19	4,46	4,54	11,10	4,63
α	0,0137	0,0135	6,12	7,69	7,94	17,30	5,17
\bar{Y}_4	0,0000	0,0000	4,55	5,83	6,00	12,30	2,87

Le tableau 4 montre que, si l'on impute toutes les valeurs des variables, l'augmentation de la valeur de m au-delà de cinq produit des gains d'efficacité appréciables. L'importance du gain dépend de la grandeur de b_m . S'il est petit relativement à \bar{v}_m , par exemple lorsqu'on impute des valeurs uniquement pour un petit nombre d'unités sélectionnées, le gain d'efficacité dû à l'augmentation de la valeur de m n'est pas important. Pour toute stratégie fondée sur des données partiellement synthétiques, les imputeurs peuvent comparer les gains d'efficacité aux compromis éventuels en ce qui concerne la confidentialité grâce à des études en simulation du comportement des intrus pour divers nombres d'ensembles de données synthétiques diffusés.

Tableau 4
Sensibilité des inférences sur données partiellement synthétiques à la valeur de m

Conditions	Var \bar{q}_m	Moy. T_p	Couv. IC à 95 %
Inférence pour β			
$m = 2$	6,52	6,50	92,7
$m = 3$	5,38	5,38	94,4
$m = 4$	4,64	4,89	95,4
$m = 5$	4,46	4,54	95,1
$m = 10$	3,87	3,88	94,4
$m = 50$	3,30	3,37	95,1
Inférence pour α			
$m = 2$	10,62	10,89	93,4
$m = 3$	8,92	9,15	94,9
$m = 4$	8,41	8,45	94,9
$m = 5$	7,69	7,94	95,4
$m = 10$	6,99	7,02	94,8
$m = 50$	6,05	6,28	95,5
Inférence pour \bar{Y}_4			
$m = 2$	8,13	7,96	93,4
$m = 3$	6,51	6,86	95,5
$m = 4$	6,11	6,33	95,6
$m = 5$	5,83	6,00	95,3
$m = 10$	5,13	5,38	95,4
$m = 50$	4,66	4,87	95,5

Les variances associées à α sont multipliées par 10^6 .

5. Conclusion

Les simulations présentées dans l'article montrent que les règles habituelles de combinaison des ensembles de données obtenus par imputation multiple peuvent produire des estimations de la variance présentant un biais par excès quand on les applique à des données partiellement synthétiques. Les nouvelles règles présentées ici semblent corriger le problème, donc produire des inférences plus

fiables. D'autres études seront nécessaires pour évaluer les propriétés de ces nouvelles règles lorsqu'on applique une stratégie basée sur des données partiellement synthétiques à des données réelles pour lesquelles les modèles d'imputation corrects sont vraisemblablement inconnus. Qui plus est, il faudrait évaluer les nouvelles règles dans le cas où les ensembles de données diffusés contiennent des imputations multiples de données manquantes, par exemple pour corriger la non-réponse partielle. Comme l'a fait remarquer un examinateur du présent article, si une part importante des imputations vise à remplacer des données manquantes, T_m pourrait ne pas donner d'aussi piètres résultats comparativement à T_p .

Les simulations et la théorie donnent aussi à penser que, dans la mesure du possible, les imputeurs devraient utiliser uniquement des données provenant des unités pour lesquelles on a choisi de remplacer certaines valeurs en vue d'estimer les lois prédictives a posteriori pour les imputations. Une étude plus approfondie de cette recommandation dans le cas de la simulation de plus d'une variable dans des ensembles de données réelles serait fort utile.

Enfin, le présent article ne traite pas de l'effet de diverses stratégies fondées sur des données partiellement synthétiques sur la protection de la confidentialité ni de la comparaison des méthodes fondées sur des données partiellement synthétiques à d'autres méthodes de contrôle de la divulgation. Ces comparaisons aideraient les imputeurs à déterminer si les premières méthodes conviennent pour leurs fichiers de microdonnées à grande diffusion.

Remerciements

La présente étude a été financée par le United States Bureau of the Census aux termes d'un contrat avec Datametrics Research. L'auteur remercie Trivellore Raghunathan, Donald Rubin et Laura Zayatz de leur appui statistique et de leurs encouragements au cours des travaux, ainsi que l'examineur et un rédacteur adjoint de leurs suggestions et commentaires constructifs.

Bibliographie

- Abowd, J.M., et Woodcock, S.D. (2001). Disclosure limitation in longitudinal linked data. Dans *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, EP. Doyle, J. Lane, L. Zayatz et J. Theeuwes (Éds.). Amsterdam: North-Holland. 215-277.
- Dandekar, R.A., Cohen, M. et Kirkendall, N. (2002a). Sensitive micro data protection using Latin hypercube sampling technique. Dans *Inference Control in Statistical Databases*, J. Domingo-Ferrer (Éd.). Berlin: Springer-Verlag. 117-125.
- Dandekar, R.A., Domingo-Ferrer, J. et Sebe, F. (2002b). LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. Dans *Inference Control in Statistical Databases*, J. Domingo-Ferrer (Éd.). Berlin: Springer-Verlag. 153-162.

- Fienberg, S.E., Makov, U.E. et Steele, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14, 485-502.
- Fienberg, S.E., Steele, R.J. et Makov, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and log-linear models. Dans *Proceedings of Bureau of Census 1996 Annual Research Conference*, 87-105.
- Franconi, L., et Stander, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician*, 51, 1-11.
- Franconi, L., et Stander, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing*. À venir.
- Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.
- Kennickell, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. Dans *Record Linkage Techniques, 1997*, W. Alvey et B. Jamerson (Éds.). Washington, D.C.: National Academy Press, 248-267.
- Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.
- Liu, F., et Little, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. Dans *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2133-2138.
- Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing*. À venir.
- Polettini, S., Franconi, L. et Stander, J. (2002). Model-based disclosure protection. Dans *Inference Control in Statistical Databases*, J. Domingo-Ferrer (Éd). Berlin: Springer-Verlag. 83-96.
- Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.
- Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18, 531-544.
- Reiter, J.P. (2003). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Rapport technique, Institute of Statistics and Decision Sciences, Duke University.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.
- Willenborg, L., et De Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.