Imputation équilibrée pour la non-réponse en fromage suisse

Yves Tillé [1] & Audrey-Anne-Vallée [1]

[1] *Institut de Statistique, Université de Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Suisse*
yves.tille@unine.ch, audrey-anne@unine.ch

Résumé. La non-réponse en fromage suisse ou la non-réponse non monotone regroupe les cas où toutes les variables d'une enquête contiennent des valeurs manquantes sans schéma particulier. L'imputation des valeurs manquantes permet de réduire le biais et la variabilité causés par la non-réponse. Il est difficile de préserver les distributions et les relations entre les variables lors de l'imputation dans le cas de non-réponse en fromage suisse. Dans cette présentation, l'imputation équilibrée par les K plus proches voisins Hasler and Tillé (2016) est développée pour le cas de la non-réponse en fromage suisse. Il s'agit d'une méthode d'imputation par donneurs qui est aléatoire et construite pour répondre à plusieurs exigences. D'abord, un non-répondant peut être imputé par des donneurs qui sont proches de lui. Les distances sont calculées avec les valeurs disponibles des variables. Ensuite, toutes les valeurs manquantes d'un non-répondant sont imputées par le même donneur choisi aléatoirement. Enfin, les donneurs sont sélectionnés de façon à ce que si on imputait les valeurs observées des non-répondants aussi, les estimations des totaux imputés et les totaux connus devraient être les mêmes. Pour imputer en respectant de telles contraintes, une matrice de probabilités d'imputation est construite à l'aide de méthodes de calage. Les donneurs sont ensuite choisis avec ces probabilités et des méthodes d'échantillonnage équilibré.

Mots-clés. imputation, fromage suisse, non-réponse

Abstract. Swiss cheese nonresponse or non monotone nonresponse occurs when all the variables of a survey can contain missing values without a particular pattern. Imputation of missing values allows to reduce the bias and the variability due to nonresponse. It is difficult to preserve the distributions and the relations between the variables when imputing in the swiss cheese nonresponse case. In this presentation, balanced K-nearest neighbour imputation Hasler and Tillé (2016) is extended to swiss cheese nonresponse. It is a donor imputation method which is random and constructed to meet some requirements. First, a nonrespondent can be imputed by donors which are close. The distances are calculated with the observed values. Next, all the missing values of a nonrespondent are imputed by the same donor. Last, the donors are chosen so that if the observed values of the nonrespondents were imputed, the estimated totals would be the same as the one calculated with the observed values. To meet all the requirements, a matrix of imputation probabilities is constructed with calibration techniques. The donors are selected with these imputation probabilities and balanced sampling methods.

Keywords. imputation, nonresponse, Swiss cheese

Introduction

La non-réponse est souvent inévitable dans les enquêtes à grande échelle. Elle peut apparaître sous deux formes. La non-réponse est dite totale lorsque les valeurs de toutes les variables d'enquête sont manquantes pour certains individus de l'échantillon. La non-réponse est dite partielle lorsqu'une partie seulement des variables est manquante pour un ensemble d'individus. Les estimations des paramètres d'intérêt peuvent être gravement affectées par les valeurs manquantes qui peuvent introduire un biais et causer une augmentation de la variabilité. La repondération des unités répondantes et l'imputation des valeurs manquantes permettent de réduire le biais et la variance causés par la non-réponse.

Dans le cas des enquêtes avec plusieurs variables d'intérêt, différents cas de non-réponse partielle peuvent se présenter. Dans un premier cas, une seule variable contient des valeurs manquantes. Les autres variables d'enquête complètement observées peuvent être utilisées pour imputer les valeurs manquantes. Un grand nombre de méthodes d'imputation ont été développées dans ce contexte. Haziza (2009) présente un survol des méthodes d'imputation déterministes et aléatoires, dont l'imputation multiple et l'imputation fractionnelle. Andridge and Little (2010) font un survol de méthodes d'imputation par donneurs. Dans un deuxième cas, la non-réponse peut affecter plusieurs variables de façon monotone. Ce schéma se prête bien au cas des études longitudinales, lorsque des individus cessent de participer aux enquêtes au cours du temps. Ce papier traite plus particulièrement d'un autre cas, la non-réponse en fromage suisse, ou la non-réponse non monotone, qui regroupe les cas où toutes les variables d'une enquête contiennent des valeurs manquantes sans schéma particulier. Les traitements pour une telle non-réponse multivariée sont peu nombreux. Il est difficile de préserver les distributions des variables et les relations entre les variables en ayant des valeurs manquantes dans tout le jeu de données. Judkins (1997) propose des méthodes d'imputation par donneur et Andridge and Little (2010) présente un petit survol des méthodes développées. Certaines méthodes permettent d'imputer de façon itérative, par exemple Raghunathan et al. (2001) utilisent une séquence de modèles de régression entre les variables.

Dans ce papier, l'imputation équilibrée par les K plus proches voisins Hasler and Tillé (2016) est développée pour le cas de non-réponse en fromage suisse. Cette méthode a plusieurs propriétés et avantages qui justifient l'intérêt de la développer pour le cas multivarié. Il s'agit d'une méthode par donneur, donc d'une méthode d'imputation aléatoire qui tend à préserver les distributions des variables. L'échantillonnage équilibré est utilisé pour diminuer la variabilité supplémentaire causée par l'imputation aléatoire. Une méthode par donneur permet aussi d'avoir un jeu de données à imputer contenant des variables continues et catégorielles. De plus, il n'y a qu'un donneur choisi pour remplacer toutes

les valeurs manquantes d'un non-répondant. Ceci permet d'assurer une cohérence entre les valeurs imputées d'un individu. Aussi, un non-répondant peut être imputé par des donneurs qui sont proches de lui, donc semblables à lui. Enfin, avec des méthodes de calage et d'échantillonnage équilibré, les donneurs sont choisis de façon à ce que si les valeurs observées étaient imputées, les estimations des totaux imputés et des totaux des valeurs connues devraient être les mêmes. Le contexte et les exigences de la méthode sont présentés dans les Sections 1 et 2. Puis la construction de la matrice de probabilités d'imputation est détaillée dans la Section 3. La sélection des donneurs est traitée dans la Section 4 et l'imputation dans la Section 5.

1 Non-réponse en fromage suisse

Considérons une population finie U de taille N avec J variables d'intérêt. Un échantillon s de taille n est sélectionné de façon aléatoire selon un plan de sondage $p(s)$. La probabilité d'inclusion d'ordre un de l'unité k est π_k et la probabilité d'inclusion d'ordre deux des unités k et ℓ est $\pi_{k\ell}$. Le vecteur de J variables d'intérêt, $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^\top$, n'est pas nécessairement complètement observé pour tout $k \in s$. Il est attendu qu'un sous-ensemble des unités échantillonnées soit complètement observé, tandis que toutes les variables du reste de l'échantillon soient sujettes à la non-réponse. Considérons donc $s_r \subset s$ un ensemble de n_r unités pour lequel les J variables sont complètement observées. Considérons aussi $s_m = s - s_r$, un ensemble de $n_m = n - n_r$ unités pour lequel des valeurs, mais pas toutes, sont manquantes. La non-réponse est non monotone, elle n'a donc pas de schéma particulier.

Supposons que les valeurs manquantes sont traitées par imputation. La valeur imputée de l'unité k à la variable j est notée x^*_{kj}. Alors le total dans la population de la variable j, $X_j = \sum_{k\in U} x_{kj}$, peut être estimé par

$$\widehat{X}_j = \sum_{k\in s} r_{kj} d_k x_{kj} + \sum_{k\in s} (1 - r_{kj}) d_k x^*_{kj},$$

où $d_k = \pi_k^{-1}$ est le poids d'échantillonnage de l'unité k et r_{kj} vaut 1 si la variable j de l'unité k est observée et 0 sinon.

2 Exigences

La méthode d'imputation proposée est élaborée pour assurer une cohérence et une précision dans le jeu de données imputé. Quatre exigences sont énoncées:

(i) Les valeurs imputées devraient être choisies parmi les valeurs des n_r unités complètement observées: une méthode par donneur doit être utilisée.

(ii) Il devrait y avoir un donneur par unité. Toutes les valeurs manquantes d'une unité devraient être imputées par le même donneur.

(iii) Les donneurs devraient être sélectionnés parmi les K plus proches voisins de l'unité avec des valeurs manquantes.

(iv) Si les valeurs observées des non-répondants étaient imputées, l'estimateur du total de toutes les valeurs observées devrait rester inchangé, et ce, pour chaque variable.

L'exigence (i) assure que les valeurs imputées sont réalistes et déjà observées, et ce, pour les variables catégorielles et continues. Aussi, avec une méthode d'imputation aléatoire, les distributions des variables tendent à être préservées. Le but de l'exigence (ii) est de préserver les relations entre les variables. L'exigence (iii) permet d'imputer un non-répondant par une unité semblable et donc avoir une cohérence entre les valeurs imputées et les valeurs observées du non-répondant. Par exemple, si le sexe et la taille de personnes sont mesurés, une taille manquante d'un homme devrait être imputée par la taille d'un homme. Puis l'idée de l'exigence (iv) est que l'information observée est inchangée si les unités avec des valeurs manquantes étaient complètement imputées. Les estimations basées sur les valeurs connues ne seraient pas affectées.

Pour implémenter une méthode d'imputation par donneur, chaque unité complètement observée reçoit une probabilité de donner ses valeurs à chacun des non-répondants. Ainsi, un donneur par non-répondant peut être choisi selon ces probabilités d'imputation. Les probabilités d'imputation respectant les exigences (i)-(iv) sont détaillées dans la Section 3. La sélection des donneurs est détaillée dans la Section 4.

3 Matrice des probabilités d'imputation

La première étape d'une méthode d'imputation par donneur est d'attribuer des probabilités d'imputation aux unités ayant complètement répondu. Considérons $\boldsymbol{\psi} = (\psi_{ik})$, où $(i,k) \in s_r \times s_m$, la matrice de probabilités d'imputation. L'élément ψ_{ik} est la probabilité que le répondant i donne ses valeurs au non-répondant k et

$$\psi_{ik} \geq 0. \tag{1}$$

Un seul donneur est choisi aléatoirement pour chaque unité $k \in s_m$ en respectant les probabilités d'imputation. Donc, la somme des probabilité d'imputation associées à un non-répondant doit être 1,

$$\sum_{i \in s_r} \psi_{ik} = 1, \tag{2}$$

pour tout $k \in s_m$. Toutes les valeurs manquantes de l'unité k sont imputées par les valeurs correspondantes du donneur. Les exigences (i) et (ii) sont alors remplies. L'exigence (iii)

limite l'ensemble de donneurs possibles d'une unité à ses K plus proches voisins. Dans ce cas, la probabilité que l'unité $i \in s_r$ donne ses valeurs au non-répondant $k \in s_m$ est non nulle seulement si i est un des K plus proches voisins de k;

$$\psi_{ik} = 0 \text{ si } i \notin \text{kpp}(k), \tag{3}$$

où $\text{kpp}(\ell) = \{j \in s_r | \, \text{rang}(d(j,\ell)) \leq K\}$ et $d(.,.)$ est une fonction de distance.

L'exigence (iv) suggère que si les valeurs observées de l'unité $k \in s_m$ étaient imputées par les valeurs correspondantes de son donneur, l'estimateur du total de chaque variable resterait le même que l'estimateur du total calculé avec les valeurs observées seulement. Les probabilités d'imputation sont donc choisies pour que si les valeurs connues des unités dans s_m étaient imputées par l'espérance de leur valeur imputée, les estimateurs du total correspondraient aux estimateurs basés sur les valeurs observées. Donc les probabilités d'imputation respectent

$$\sum_{k \in s_m} d_k r_{kj} \sum_{i \in s_r} \psi_{ik} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}, \tag{4}$$

pour $j = 1, \ldots, J$.

L'équation (4) peut être réécrite comme

$$\sum_{i \in s_r} \left(\sum_{k \in s_m} d_k r_{kj} \psi_{ik} \right) r_{ij} x_{ij} = \sum_{k \in s_m} d_k r_{kj} x_{kj}, \tag{5}$$

pour $j = 1, \cdots, J$. Les probabilités d'imputation respectant (5) peuvent être trouvées par calage Deville and Särndal (1992). Considérons les probabilités d'imputation initiales

$$\psi_{ik}^0 = \begin{cases} \dfrac{1}{K} & \text{si } i \in \text{kpp}(k), \\ 0 & \text{sinon.} \end{cases} \tag{6}$$

Des probabilités d'imputation finales ψ_{ik} proches de ψ_{ik}^0 respectant (5) sont cherchées. Considérons la fonction de distance

$$G(\psi_{ik}, \psi_{ik}^0) = \psi_{ik} \log(\frac{\psi_{ik}}{\psi_{ik}^0}) + \psi_{ik}^0 - \psi_{ik},$$

alors les ψ_{ik} sont obtenus en minimisant

$$\mathcal{L} = \sum_{k \in s_m} \sum_{i \in s_r} G(\psi_{ik}, \psi_{ik}^0) - \sum_{j=i}^{J} \lambda_j \left[\sum_{i \in s_r} \sum_{k \in s_m} d_k r_{kj} \psi_{ik} r_{ij} x_{ij} - \sum_{k \in s_m} d_k r_{kj} x_{kj} \right].$$

En utilisant

$$\frac{\partial \mathcal{L}}{\partial \psi_{ik}} = \log \frac{\psi_{ik}}{\psi_{ik}^0} - \sum_{j=1}^{J} \lambda_j d_k r_{kj} r_{ij} x_{ij} = 0,$$

les probabilités d'imputation sont

$$\psi_{ik} = \psi_{ik}^0 \exp\left[\sum_{j=1}^{J} \lambda_j d_k r_{kj} r_{ij} x_{ij}\right]. \tag{7}$$

Des techniques de calage sont utilisées pour trouver $\lambda = (\lambda_1, \ldots, \lambda_J)^\top$ respectant les Équations (1)-(4).

4 Matrice d'imputation

Une fois la matrice de probabilités d'imputation $\boldsymbol{\psi}$ complétée, les donneurs peuvent être choisis de façon aléatoire. Considérons $\boldsymbol{\phi} = (\phi_{ik})$, où $(i,k) \in s_r \times s_m$, la matrice d'imputation. L'élément ϕ_{ik} vaut 1 si l'unité i est choisie pour donner ses valeurs à l'unité k, 0 sinon. Un seul donneur est choisi par non-répondant, donc

$$\sum_{i \in s_r} \phi_{ik} = 1.$$

Pour respecter l'exigence (iv), les donneurs doivent être choisis de façon à respecter

$$\sum_{k \in s_m} \sum_{i \in s_r} \frac{\phi_{ik}}{\psi_{ik}} d_k r_{kj} \psi_{ik} x_{ij} = \sum_{k \in s_m} \sum_{i \in s_r} d_k r_{kj} \psi_{ik} x_{ij}. \tag{8}$$

L'échantillonnage équilibré Deville and Tillé (2004) est utilisé pour respecter les contraintes d'équilibrage (8). Afin d'assurer qu'un seul donneur est choisi par non-répondant et que les contraintes d'équilibrage sont respectées une fois les donneurs choisis, la matrice $\boldsymbol{\phi}$ est générée avec l'échantillonnage stratifié équilibré Chauvet (2009); Hasler and Tillé (2014). Un total de n_m strates sont formées et un donneur est choisi par strate. Une strate correspond à un non-répondant. La probabilité d'inclusion utilisée dans l'échantillonnage stratifié équilibré est ψ_{ik} et sa variable d'équilibrage associée est $d_k r_{kj} \psi_{ik} x_{ij}$, pour $(i,k) \in s_r \times s_m$.

5 Imputation et discussion

L'imputation du jeu de données se fait à partir de la matrice $\boldsymbol{\phi}$. La valeur manquante de l'unité $k \in s_m$ pour j tel que $r_{kj} = 0$ est imputée par

$$x_{kj}^* = \sum_{i \in s_r} \phi_{ik} x_{ij}.$$

Les exigences (i)-(iii) sont alors parfaitement respectées. L'exigence (iv) est parfaitement ou approximativement respectée, selon les limites des méthodes d'équilibrage.

Il est aussi possible d'utiliser une version déterministe de la méthode d'imputation proposée. Il suffit d'utiliser l'espérance de ϕ_{ik} pour $(i,k) \in s_r \times s_m$, soit ψ_{ik}. Ainsi, la valeur manquante de l'unité $k \in s_m$ pour j tel que $r_{kj} = 0$ est imputée par

$$x^*_{kj} = \sum_{i \in s_r} \psi_{ik} x_{ij}.$$

Il ne s'agit alors plus d'une méthode d'imputation par donneur. Toutefois, l'exigence (iv) est parfaitement respectée.

Les méthodes d'imputation aléatoires causent généralement une augmentation de la variance des estimateurs imputés. L'imputation équilibrée permet de minimiser, ou même d'éliminer, cette variabilité supplémentaire. Dans les travaux à venir, un estimateur de variance devrait être développé et une étude par simulations devrait permettre d'analyser les propriétés théoriques d'estimateurs imputés et de leur variance.

References

Andridge, R. R. and Little, R. J. A. (2010). A review of dot deck imputation for survey non-response. *International Statistical Review*, 78:40–64.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35:115–119.

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

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

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

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

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C. R., editors, *Sample surveys: Design, methods and applications*, pages 215–246. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium*, page 97, Statistics Canada.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95.