Comparaison des différentes méthodes d'imputation de la non-réponse de lien dans l'enquête SD2001

Carme Caum Julio[1]: carme.caum@gmail.com

Encadrant : **Maryse Marpsat**[2]**, Monique Bécue**[1].

Contexte, problématique et objectifs :

L'enquête nationale sans-domicile (SD) menée par l'Insee et l'Ined, en raison de sa méthode particulière d'échantillonnage indirect (tirage des prestations au lieu d'individus), nécessite de pondérer les données par la méthode de partage des poids (Lavallée, 1995) pour réaliser des estimations au sens des individus. Un repérage du nombre de liens hebdomadaires entre la population d'usagers et celle des prestations a été collecté sous forme de **semainier** dans le questionnaire (l'enquêté étant interrogé sur où il avait dormi et mangé les sept jours précédents au jour de l'entretien). Compte tenue que la présence de liens manquants entraîne des surestimations sur un total d'intérêt (Xu Xiaojian, Lavallée 2009), nos objectifs sont les suivants :

- Valider la méthode d'imputation du semainier pour l'enquête de 2001 et proposer des variantes.
- Estimer un modèle Bayésien pour expliquer et prédire le nombre hebdomadaire de liens (ou prestations).
- Développer un plan d'action pour l'imputation du semainier de l'édition de 2012.

Méthodologie

Étant donné l'intérêt de valider la méthode d'imputation de 2001, notre démarche a été la suivante: des données manquantes ont été générées sur l'**ensemble des répondants** (à l'aide de modèles de mesures répétées estimés sur les non-répondants ou NR); deux familles de méthodes d'imputation on été testées sur l'ensemble de NR générés et, finalement, l'erreur relatif (ER) dans l'estimation du nombre de SDF a été prise comme critère de validation (Xu Xiaojian, Lavallée 2009).

Les deux familles de méthodes d'imputation explorées sont: **imputation par plus proche voisin** et imputation à l'aide de modèles Bayésiens. Pour la première famille trois paramètres doivent être fixés : la strate/s pour la quête de possibles donneurs, le critère de sélection (mesurée entre un NR et un donneur candidat) et les **modalités du semainier (détaillé ou binaire)**. En modifiant ce dernier paramètre, nous avons réalisé 40 simulations de la méthode employée en l'édition de 2001 (semainier détaillé, SD 2001) et 40 simulations de la méthode avec le semainier binaire (SD links). Pour la deuxième famille, un **modèle Bayésien à deux niveaux dépendants** a été estimé pour prédire le nombre de liens hebdomadaires par type de lien (dormir, repas midi, repas soir). Dix-huit variables explicatives on été introduites pour l'estimation du modèle complet, avec des distributions a priori non-informatives. Les modèles, au-delà d'être employées comme méthode d'imputation, ont servi à proposer de strates intéressantes pour les méthodes d'imputation par plus proche voisin.

Résultats et Conclusions

1. La simplification du semainier (au sens des types de services) avec la méthode links apparaît, d'auprès les simulations, comme une amélioration au moment de choisir l'ensemble de donneurs pour un NR.
2. Les modèles Bayésiens présentent des avantages et des résultats assez bons (médiane de ER plus proche à 0). Néanmoins, la variance de l'ER demeure trop grande (principalement due à la complexité du modèle). Nous proposons ré-estimer les modèles avec la méthode itérative forward afin de trouver des modèles plus simples.
3. Notre proposition pour SD 2012 est la suivante:

✓ Préconiser l'application de la méthode de donneurs dite SD links sous un taux de non-réponse de lien jusqu'à 5%, en choisissant comme strates: le type de service où l'enquête a eu lieu et/ou le fait d'être en couple ou pas.

✓ Pour les questionnaires auto-administrés (30% NR au pilote de 2011 et plusieurs NR totaux), appliquer la méthode des modèles Bayésiens mis à jour en introduisant l'information de l'enquête de 2001 comme information *a priori*. Ceci fournira des prédictions plus précises.

[1] Faculté de Mathématiques et Statistiques de l'Universitat Politécnica de Catalunya (UPC).

[2] Insee, Ined.