Application de la méthode d'imputation "SwissCheese" dans l'enquête sur les revenus et les conditions de vie (SILC)

Michael Leuenberger

Office fédérale de la statistique OFS
Science des données, IA et méthodes statistiques | Méthodes statistiques

Journées suisses de la statistique publique | 26-27 octobre 2022

Introduction

Un projet en collaboration avec l'*institut de statistique de l'université de Neuchâtel*. Un remerciement particulier à *Esther Eustache* et *Arnaud Tripet*.

Objectif principal :
Mettre en évidence la procédure d'utilisation et les adaptations apportées à l'algorithme pour une enquête spécifique :

- Données *SILC* : enquête sur les revenus et les conditions de vie.
  - 7 variables d'intérêt du module fortune.
- Algorithme *SwissCheese* : algorithme d'imputation basé sur l'équilibrage des totaux.

Les données SILC

- En raison des filtres appliqués à chaque variable d'intérêt de l'enquête SILC 2020, 7 ensembles de données ont été extraits.
- Les quantités de données disponibles et les taux de valeurs manquantes peuvent varier fortement d'un ensemble à l'autre (de 9% à 19%).
- L'imputation simultanée de l'ensemble des données n'est donc pas recommandée.

Les variables SILC

Cette présentation se concentrera sur :

- *HF5050* : Fortune - possession d'objets de valeur : montant.
  - Nombre d'observations : 3048
  - Nombre de valeurs manquantes : 338 (11%)
- *HV070* : Dettes - total des hypothèques sur résidence principale : montant.
  - Nombre d'observations : 2995
  - Nombre de valeurs manquantes : 311 (10%)
- Un total de 127 variables auxiliaires sont utilisées (telles que des informations sur la composition du ménage et la situation financière).

L'algorithme SwissCheese

Développé par l'université de Neuchâtel, cet algorithme peut gérer l'imputation multivariée et présente les propriétés suivantes :

- Les valeurs manquantes sont imputées par des valeurs réelles en sélectionnant un donneur parmi les répondants dans le voisinage du receveur.
- Les relations entre les variables imputées sont préservées.

Le choix du donneur est basé sur un compromis entre la définition du voisinage et l'équilibrage des totaux des variables auxiliaires.

Cadre de simulation

À partir d'une première étude préliminaire basée sur les données de l'enquête SILC 2015, le processus a été adapté au jeu de données SILC 2020. Il se compose des étapes suivantes :

- Détermination de groupes de réponses homogènes à partir des feuilles terminales d'un modèle d'arbre de régression.
- Génération de valeurs manquantes dans chaque groupe homogène sur la base du taux de manquants observé.
- Sélection de variables auxiliaires en fonction de leur corrélation avec la variable d'intérêt.
- Application de l'algorithme SwissCheese sur les données simulées.

Évaluation

Les comparaisons avec l'algorithme MissForest ont été effectuées avec les outils et mesures suivants :

- Matrice de confusion basée sur les quintiles.
- Boxplots basés sur les déciles.
- Mesure de la précision (ACC) basée sur la matrice de confusion et l'erreur quadratique moyenne (RMSE) basée sur les valeurs originales.

Résultats

Jeu de données	ACC SwissCheese	ACC MissForest	RMSE SwissCheese	RMSE MissForest
HV070	45.3%	61.7%	0.61	0.38
HF5050	43.4%	46.7%	1.01	0.83

Les cinq autres variables ont des résultats similaires pour la plupart.

Résultats - HF5050

Valeurs imputées	Valeurs originales Q1	Q2	Q3	Q4	Q5
Q1	0.60	0.20	0.15	0.09	0.02
Q2	0.04	0.07	0.15	0.01	0.02
Q3	0.16	0.41	0.40	0.16	0.02
Q4	0.13	0.29	0.30	0.42	0.38
Q5	0.07	0.02	0.00	0.31	0.57
Tot	1.00	1.00	1.00	1.00	1.00

Matrice de confusion du jeu de données HF5050 avec les imputations issues de l'algorithme SwissCheese.

Résultats - HV070

Valeurs imputées \ Valeurs originales	Q1	Q2	Q3	Q4	Q5
Q1	0.71	0.38	0.08	0.10	0.00
Q2	0.05	0.19	0.15	0.00	0.09
Q3	0.14	0.31	0.38	0.25	0.17
Q4	0.05	0.12	0.27	0.35	0.17
Q5	0.05	0.00	0.12	0.30	0.57
Tot	1.00	1.00	1.00	1.00	1.00

Matrice de confusion du jeu de données HV070 avec les imputations issues de l'algorithme SwissCheese.

Conclusions

- Résultats encourageants de l'algorithme SwissCheese.
- Aux extrêmes, l'algorithme SwissCheese obtient de meilleurs résultats que l'algorithme MissForest.
- Amélioration importante des résultats de l'imputation avec l'ajout de variables fourchettes (des variables d'intérêt) en tant que variables auxiliaires.

Améliorations potentielles :

- Tester d'autres procédures de sélection de variables.
- Ajouter des variables auxiliaires supplémentaires.

Références

- E. Eustache, A.-A. Vallée and Y. Tillé. Balanced Donor Imputation Handling Swiss Cheese Nonresponse. *Accepted paper in Statistica Sinica*.
- Daniel J. Stekhoven and P. Bühlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1) :112-118, 2012.

Package SwissCheese disponible sur :
https ://github.com/EstherEustache/SwissCheese