Imputation de l'Enquête Bugdet de Famille 2000

Nicolas CHOPIN, INSEE
Emmanuel MASSE, MEDD
Céline ROUQUETTE, INSEE

Présentation de l'Enquête

- Réalisée tous les cinq ans
- Vise à reconstituer la comptabilité du ménage: ressources et dépenses (y compris dépenses non liées à la consommation: impôts et taxes, remboursement de crédits, etc.)
- Point de départ de nombreuses études (déterminants de la consommation, niveaux de vie, inégalités, etc.)

Structure de l'Enquête

Questionnaires

- Revenus
- caractéristiques socio-démographiques
- Dépenses importantes (logement, biens d'équipement, etc.)

Carnets

- Dépenses faibles ou fréquentes
- Période de 14 jours
- Rempli par chaque membre du ménage de plus de 14 ans

Non-réponse dans l'enquête BDF 2000

- L'enquête est affectée de non-réponse partielle et totale.

Non-réponse totale

Absence totale d'information sur une unité donnée (traitée par calage)

Non-réponse partielle

Absence d'information limitée à une certaine variable une unité donnée (traitée par **imputation**)

Quelles valeurs imputer?

Les techniques d'imputation doivent idéalement respecter:

- la structure **globale** de la variable imputée (moyenne, variance)
- La cohérence au niveau « **individuel** » (relation entre les variables)

Imputation des revenus

Méthodes des « résidus simulés »

Soit Y le log d'une variable de revenus (salaire, épargne, etc.) et X différentes caractéristiques du ménage, on suppose:

$$Y = Xb + sU$$

Puis:

1) On estime b et s (estimateurs b_est et s_est)
2) On remplace chaque valeur manquante Y_i par:X_i*b_est+s_est*U_i, où U_i est une v.a. N(0,1)

Avantages et inconvénients d'une méthode économétrique

Avantages

- Permet de restituer simplement les relations entre les variables
- Redressement du biais dû à la non-indépendance entre Y (variable à imputer) et R (indicatrice de non-réponse)

Inconvénients

- Méthode peu automatisable
- Impose un modèle (normatif)

Imputation des dépenses du questionnaire

Plus de 130 variables de montant de dépense sont affectées de non-réponse. Nous avons eu recours à:

- Une imputation économétrique, dès que le nombre de montants observés sont suffisants et l'ajustement du modèle (R^2) est satisfaisant
- Une imputation par hot-deck stratifié, dans tous les autres cas

Hot-Deck stratifié pour l'imputation des dépenses

- Construction de dix strates, à partir des déciles du revenu global (épargne non incluse).
- Dans une strate donnée, imputation d'une valeur tirée aléatoirement par les montants correctement déclarés.

Imputation des carnets

Un fichier de 1110284 enregistrements pour 10300 ménages.

Le codage s'effectue dans une nomenclature à 6 positions.

Des taux de non-réponse faibles :

- 6,4 % des libellés ne sont pas complètement codés
- 1,9 % des libellés sont codés sur moins de 4 positions

Imputations des carnets

Libellé Produit	Code Produit	Libellé magasin	Code magasin	Montant
Fruits et légumes	011***	Intermarché	1112	18,10
Course	******	Auchan	1111	1226,76
EDF GDF (facture)	045***	EDF	6711	1960,00
Fruits et légumes	011***	Champion	1112	14,60
Fruits et légumes	011***	Champion	1112	13,90
Fruits et légumes	011***	Champion	1112	8,20
Livre revue	095***	Tabac	2224	6,00

Les étapes de l'imputation

- Codage automatique (SICORE)
- Traitements manuels

- Traitements semi-automatiques
- Codage sur 4 positions
- Codage sur 6 positions

Objectifs de l'imputation

- Cohérence macroéconomique des données
- Cohérence microéconomique des données

Traitements semi-automatiques

- Objectif : exploiter au maximum l'information contenue dans le libellé

- Exemple : Fruits et légumes
- Allocation aléatoire de l'ensemble de la dépense au code « fruits » ou au code « légumes »
- Le tirage se fait proportionnellement aux dépenses relatives de consommation en fruits et légumes des répondants.

Codage sur 4 positions

- On distingue deux types de libellés à imputer :
  - Ceux relatifs à un enregistrement isolé (exemples : un ticket de caisse, un achat isolé, une facture EDF-GDF, …)
  - Ceux relatifs à une dépense parmi d'autres (le libellé « crémerie » sur un ticket de caisse, …)
- On fait l'hypothèse que les premiers correspondent à des dépenses éventuellement multiples
- Alors que les seconds sont des dépenses isolées

Codage sur 4 positions

- Application de la méthode du plus proche donneur

- Utilisation des variables :
  - catégorie de commune (communes de plus ou de moins de 200000 habitants)
  - tranche d'âge de la personne de référence (moins de 35 ans, de 35 à 60 ans, plus de 60 ans)
  - code produit de la partie renseignée
  - code magasin de la partie renseignée
  - Montant déclaré de la dépense

Codage sur 4 positions

- Dans le cas de l'imputation d'un ticket (libellé isolé), on constitue à partir du fichier une table de tickets par agrégation des libellés. Et, on utilise comme variables pour la fonction distance :
  - Le type de commune, l'âge, le code magasin
  - Le code produit majoritaire en montant pour le donneur
  - Le montant des dépenses agrégées pour le donneur
- Dans le cadre de l'imputation des libellés isolés, à un enregistrement, on peut imputer plusieurs dépenses (exemple : ticket de course)

Codage à 6 positions

- Le passage de 4 à 6 positions se fait par attribution aléatoire d'un des codes produits
- Proportionnellement à la consommation relative des répondants dans les différents postes de la nomenclature

Bilan

	Total
Enregistrements (fichier initial)	1 150 614
Enregistrements (parmi les répondants)	1 110 284
Enregistrements à imputer	20 336
Enregistrements à imputer (après traitement semi-automatique)	14 871
dont enregistrements isolés à imputer (tickets) : type 2	5 627
dont enregistrements au sein d'un ticket à imputer : type 1	9 244
Nombre final d'enregistrements	1 133 408

Conclusion

- Une enquête qui nécessite l'utilisation d'une grande variété de méthodes d'imputation
- La mise en œuvre de techniques évoluées (régression économétrique, hotdeck stratifié, méthode des donneurs)
- Un très grand nombre de variables à imputer