COMPARAISON DE QUATRE MÉTHODES D'IMPUTATION DES REVENUS MOBILIERS DANS LE CADRE DE L'ENQUÊTE EU-SILC[1]

Modou DIA[2] (*)
(*) CEPS-INSTEAD Esch sur Alzette au Luxembourg[3]

Introduction

L'EU-SILC collecte des données relatives aux revenus et aux conditions de vie des ménages privés. L'étendue des revenus recouvre un large spectre : salaires, indemnités de chômage, pensions, revenus d'indépendants ou de professions libérales, revenus de transfert, mais aussi des revenus immobiliers et des revenus mobiliers. Toutes ces composantes sont sujettes à des taux de non-réponses partielles très variables. Dans le souci de réduire au maximum d'éventuels biais dans le calcul des revenus agrégés et des indicateurs en découlant, il s'avère nécessaire d'imputer ces non-réponses partielles. C'est l'imputation de ces non-réponses partielles pour les revenus mobiliers, agrégés dans une seule variable dans le questionnaire-ménage, qui est l'objet de ce document avec le plan suivant :

- d'abord exposer la problématique de l'imputation des revenus mobiliers ;
- ensuite énoncer les considérations générales sur les données utilisées ;
- puis décliner les méthodes d'estimation sollicitées aussi bien dans leur conception que dans leur mise en œuvre ;
- puis valider par un certain nombre de tests les résultats des différentes méthodes d'estimation appliquées
- enfin tenter en conclusion de désigner la méthode d'estimation la moins mauvaise au crible de ces tests.

1. Problématique

L'implémentation d'un processus d'imputation n'est pas un exercice exempt de risques et d'obstacles. Les niveaux de difficulté dépendent de la nature de la composante du revenu. Ainsi est-il plus aisé d'estimer les composantes salariales en disposant de paramètres "objectifs", entre autres, tels que le niveau d'éducation ou de formation, l'expérience professionnelle, le secteur d'activité ainsi qu'une source externe fiable permettant de tester avec une certaine consistance la validité de l'estimation réalisée. Encore mieux pour l'imputation de certains types d'allocation, un barème disponible permet de calculer de manière déterministe le montant des allocations perçues en fonction de la composition du ménage. Par contre, toute autre est l'ampleur des difficultés dès qu'il s'agit de traiter de l'imputation des revenus mobiliers. Ces difficultés sont tributaires de certaines caractéristiques propres à cette composante de revenu.

«*…Durant l'année 2007, vous-même (ou un membre de votre ménage) avez-vous bénéficié d'intérêts d'épargne, de dividendes ou de bénéfices tirés d'investissements en capital ? … Si oui, Pourriez-vous nous dire le montant annuel ?* » D'abord c'est ainsi qu'est posée la question relative aux revenus mobiliers dans l'enquête. Il en ressort qu'ils sont ainsi composés d'un portefeuille de produits très divers avec une complexité et une volatilité très variables. Ensuite, le taux de données manquantes

[1] European Union-Survey on Income and Living Conditions

[2] Je remercie mon collègue Jean-Yves Bienvenue pour ses remarques et ses observations. Pour autant j'assume personnellement l'entière responsabilité d'éventuelles erreurs qui subsisteraient.

[3] Centre d'Études de Populations, de Pauvreté et de Politiques Socio-Economiques / International Network for Studies in Technology, Environment, Alternatives, Development
3, avenue de la Fonte L-4364 Esch-sur-Alzette, Grand-duché du Luxembourg.
Adresse électronique : Modou.Dia@ceps.lu Tél. : +352 58 58 55 544 Fax : +352 58 55 714
Site web **:** **http://www.ceps.lu/**

est relativement élevé autour de 36-38% (Voir tableau 1 ci-dessous). A ces deux éléments, il faudrait ajouter l'absence de paramètres "objectifs" pour l'élaboration d'un modèle d'imputation à la différence de certaines composantes mentionnées ci-dessus. Enfin, il n'existe pas de source externe exhaustive et fiable pour évaluer la validité d'un modèle adopté.

Tableau 1 : Statistiques sur les données manquantes ou observées des revenus mobiliers

Indicateurs \ FLAGS	Revenus Mobiliers observés	Revenus Mobiliers manquants	Total
Fréquences non pondérées	1214	740	1954*
Fréquences pondérées	70449	40917	111365**
Pourcentages non pondéres	62,13%	37,87%	100%
Pourcentages pondéres	63,26%	36,74%	100%

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

**: sur 3 779 ménages répondants à l'enquête, soit un taux de possession de revenus mobiliers de l'ordre de 52%.*
***: soit un taux estimé de possession de revenus mobiliers de l'ordre de 59% sur la population des ménages.*

2. Considérations générales sur les données

C'est dans ce sous-chapitre que vont être abordés la question de l'opportunité de la pondération, le choix des variables explicatives ainsi que le découpage éventuel des variables continues élues.

2.1. Faut-il pondérer ou non ?

La source de données est constituée des données de la vague d'enquête EU-SILC collectées en 2008 et qui comporte 3 779 observations de ménages privés au sein desquelles des informations sur les revenus mobiliers ont été recueillies. A l'échantillon initial de la première vague en 2003 sur le terrain, il est ajouté chaque année un échantillon supplémentaire pour corriger les éventuels biais causés par le phénomène de l'attrition. Chaque échantillon supplémentaire est prélevé parmi la population, ayant migré au Luxembourg depuis la vague précédente, composée majoritairement de ménages « à profils jeunes ». L'ajout annuel d'échantillons supplémentaires peut vraisemblablement engendrer à terme une modification significative de la distribution de certaines variables pertinentes ou relatives à la perception de revenus mobiliers dans l'échantillon global par rapport à la distribution de ces mêmes variables au sein de la population de référence. Pour vérifier cette hypothèse, certaines variables sont sollicitées pour calculer et comparer leur distribution non pondérée dans l'échantillon[4] et celle pondérée dans la population. Ces variables sont les suivantes: l'âge du chef de ménage, le fait de percevoir ou non des revenus et le statut d'occupation du logement du ménage. Ainsi comme on peut le constater dans le tableau 2 ci-dessous entre les données collectées en 2008 et les données collectées en 2010, la distribution [pondérée] de ces variables au niveau de la population est plus stable que la distribution équivalente [non pondérée] au niveau de l'échantillon. Ainsi une économie de temps pourra être réalisée dans la mesure où le modèle retenu in fine pourra être répliqué dans ses grandes lignes dans les vagues suivantes en faisant appel aux poids dans la construction des modèles.

Il en résulte que toute modélisation de revenus mobiliers qui ne prendrait pas en compte la surreprésentation ou la sous-représentation de catégories défavorisées dans l'accès et dans la répartition de ces revenus risque d'aboutir respectivement à des erreurs de sous-estimation globale ou à des à des erreurs de surestimation globale. Cet argument constitue une raison supplémentaire pour conforter dans toutes les étapes où cela est envisageable : le choix de la démarche «pondérée» sens propre et au sens figuré pour éviter ou corriger ces éventuelles erreurs. Il s'agit d'un choix logique en se basant sur la constance de la distribution pondérée à travers les vagues d'un ensemble de variables typiques de la perception d'un revenu mobilier.

[4] En fait, il y a un abus de langage dans l'utilisation du concept «échantillon» dans ce cas précis car il fait plutôt référence dans le cas d'espèce à l'ensemble des répondants. Et les variables de contrôle utilisées ne sont pas du tout renseignées dans la base de sondage ou bien le sont dans l'unité d'échantillonnage qu'est le ménage fiscal. Or ce niveau du ménage fiscal est différent de l'unité de collecte et de l'unité d'analyse constituées par le ménage-logement. Autrement dit, il s'agit d'un sondage indirect.

2.2. Choix des variables et découpage éventuel des variables continues

Les variables retenues dans les modèles sont extraites d'un ensemble de variables dont la liste est reproduite in extenso dans l'annexe à la fin du texte. Elle comprend des variables aussi bien de types catégoriel que continu. Les variables continues sont construites pour les besoins de la modélisation et sont au nombre de trois :

- Le revenu total[5] du ménage amputé de ses composantes mobilières ;
- Le loyer générique[6];
- La taille en nombre d'équivalents du ménage d'après l'échelle modifiée d'Oxford[7].

Deux outils seront employés pour déterminer les variables incluses dans les modèles : la matrice de corrélations et la classification automatique des variables.

Le critère recommandé dans l'exploitation des matrices de corrélations consiste à choisir les variables les plus corrélées avec la variable dépendante, mais en même temps les moins corrélées entre elles. Ainsi, on constate d'abord dans le tableau 3 ci-dessous que les variables les plus corrélées avec les revenus mobiliers (en couleur rouge) sont dans l'ordre décroissant : le revenu total, le loyer générique, HS120 (*"Vivez-vous facilement avec vos revenus ?"*), HS140 (*"Vos coûts du logement sont-ils une charge lourde")* et l'âge du chef de ménage. Ensuite, Il est à mentionner deux corrélations relativement élevées d'une part entre le revenu total et le loyer générique (en couleur bleue), et d'autre part entre le revenu total et la variable HS120 (en couleur violette). Puis la variable « le loyer générique » et la variable HS120 sont exclues des variables explicatives du fait de cette corrélation forte avec la variable la plus corrélée aux revenus mobiliers c'est-à-dire le revenu total. Enfin de compte d'après l'examen de la matrice de corrélations, les variables retenues comme variables explicatives sont par élimination : le revenu total, HS140 (*"Vos coûts du logement sont-ils une charge lourde")* et l'âge du chef de ménage.

Quant aux critères des classifications automatiques, ils sont exprimés en deux versions :

- a) version 1 : une classification avec la méthode centroïde
- b) version 2 : une classification ayant comme contrainte le nombre maximal de groupes égal à quatre, cette version donne des résultats identiques à la configuration par défaut de la procédure de classification du logiciel SAS où la combinaison linéaire se base sur la première composante principale.

Dans la classification avec la méthode centroïde illustrée dans le graphique 1 ci-dessous, les revenus mobiliers participent au même groupe que les variables HS120 et HS140. Alors que dans le graphique 2 ci-dessous avec la classification avec un nombre maximal de groupe égal à quatre, les revenus mobiliers partagent le même groupe que les variables : le loyer générique, le revenu total, la source de revenu d'indépendant. Dans les deux versions ci-dessus, les deux variables «revenu total brut» et loyer générique se trouvent dans le même groupe confirmant ainsi la corrélation existante dans la matrice des corrélations ci-dessus. Dans la version 2 (voir graphique 2), les variables HS120 et HS140 sont assemblées dans un même groupe comme illustration de leur forte corrélation. En réunissant toutes variables cohabitantes de groupe que la variable «revenus mobiliers» et en éliminant une variable parmi les binômes fortement corrélés, on aboutit à cet ensemble de variables explicatives finales par le biais de la classification : le revenu total, HS140, la source de revenu d'indépendant. Pour la solution finale relative aux variables explicatives retenues, il faut faire la synthèse de ce résultat issu de l'outil de classification avec celui provenant de l'outil de la matrice de corrélation qui est ce suivant : le revenu total, HS140, l'âge du chef de ménage. Le choix final des variables explicatives va comprendre outre l'intersection de deux ensembles constituée des deux variables «le revenu total» et «HS140», mais aussi «l'âge du chef de ménage» par choix raisonné en

[5] Par suite sauf indications contraires, par revenu total brut du ménage, cela s'entend hors revenus mobiliers.
[6] Par loyer générique, il faudrait comprendre le loyer intégrant le loyer fictif ou imputé pour les non-locataires de même que le loyer réel payé par les locataires au prix du marché ou au dessous du prix du marché.
[7] Selon cette échelle, la valeur «1» est attribuée à la première personne adulte, la valeur «0.5» est affectée aux autres adultes et la valeur «0.3» est assignée aux enfants c'est-à-dire aux personnes âgées de moins de quatorze ans

présumant son pouvoir explicatif plus grand au détriment de la variable «la source de revenu d'indépendant ». Par ailleurs, cette dernière variable pourrait être aussi récusée pour sa moins grande fiabilité d'autant plus que, entre autres, des salariés peuvent être de «vrais-faux indépendants» en étant légalement obligés de se déclarer comme tels pour la commercialisation d'une œuvre intellectuelle (livres, films etc.). En fin de compte, les trois variables dites indépendantes des modèles prédictifs à venir sont : «le revenu total», «HS140», «l'âge du ménage». La limite fixée à un nombre de trois variables pourrait se justifier par une interprétation restrictive du principe de parcimonie des paramètres.

Ces trois variables explicatives seront utilisées dans la mise en œuvre des quatre méthodes d'imputation avec le même découpage dans le souci d'une meilleure comparabilité entre les méthodes durant la phase de validation. Les autres variables fortement corrélées aux revenus mobiliers et non sélectionnées seront mises à contribution précisément à cette étape de validation.

Pour toutes les méthodes d'estimation envisagées, chaque observation du fichier des ménages répondants doit impérativement faire partie d'une classe d'imputation pour que les données manquantes éventuellement y associées puissent être imputées le cas échéant. Pour ce faire, toutes les variables explicatives doivent être renseignées pour toutes les observations. C'est la raison pour laquelle les données manquantes de ces variables ont été imputées. Mais l'âge du chef de ménage et le revenu total étant renseignés pour toutes les observations, seule la variable catégorielle HS140 possède des observations manquantes. Ces dernières ont été imputées par le modèle logistique généralisé du logiciel d'imputation multiple IVEware (Imputation Variance Estimation ware) de l'Université de Michigan (http://www.isr.umich.edu/src/smp/ive/).

Une fois toutes les variables explicatives complètement renseignées par collecte ou par imputation, le découpage des variables éventuellement continues reste le seul obstacle à la constitution des classes d'imputation par croisement des variables explicatives catégorielles. Sur ce plan, la variable HS140 étant déjà une variable catégorielle à l'origine avec trois modalités, ce sont les variables «Age du chef de ménage» et «Revenu total» qui doivent faire l'objet d'un découpage. Ainsi l'âge du chef de ménage a été découpé en quatre modalités en tenant compte du cycle de vie et de l'effectif de chaque modalité comme dans le tableau 2 ci-dessous. Quant au revenu total, il est scindé en quintiles. Le nombre total de classes s'élève à 3 (HS140) * 4 (âge du chef de ménage) * 5 (revenu total), 60 classes pour environ 1200 observations avec des données originales des revenus mobiliers comme base du processus d'imputation[8]. Il est important souligner qu'en remplaçant les variables continues par leurs valeurs dichotomisées, les corrélations déjà calculées ci-dessus ne changent pas sensiblement.

[8] La distribution des observations donneuses et des observations receveuses ainsi que leur ratio sont consultables dans l'annexe ci-dessous. Dans le compromis entre la finesse et la consistance du découpage des cellules, le primat est donné à la finesse vu les limites inhérentes à la taille de l'échantillon.

Tableau 2 : Distribution de variables relatives à la possession de revenus mobiliers ou de variables influentes sur les revenus mobiliers

Variable	Modalité	Vague d'enquête	Pourcentage non pondéré	Pourcentage pondéré
CLASSE D'AGES	16_34 ans	6	23,3	15,5
		7	18,5	15,4
		8	15,4	15,0
	35_49 ans	6	36,0	34,7
		7	35,8	34,9
		8	35,6	35,1
	50_64 ans	6	24,7	26,5
		7	27,2	27,5
		8	29,7	28,0
	65 ans ou +	6	16,0	23,3
		7	18,5	22,2
		8	19,3	21,9
REVENUS MOBILIERS	NON	6	48,3	41,0
		7	43,8	40,9
		8	41,4	40,5
	OUI	6	51,7	59,0
		7	56,2	59,1
		8	58,6	59,5
STATUT D'OCCUPATION	Locataire avec loyer < prix du marché	6	4,0	4,8
		7	4,0	5,2
		8	3,6	3,3
	Locataire avec loyer au prix du marché	6	34,5	21,4
		7	27,3	24,8
		8	23,1	30,4
	Occupant (e) à titre gratuit	6	3,4	3,3
		7	3,0	3,2
		8	2,4	1,9
	Proprietaire	6	58,2	70,5
		7	65,8	66,8
		8	70,9	64,3

Source Enquête EU-SILC 2006-2008, CEPS/INSTEAD-STATEC

Tableau 3 : Matrice de corrélations des variables candidates
au statut de variables explicatives et des revenus mobiliers

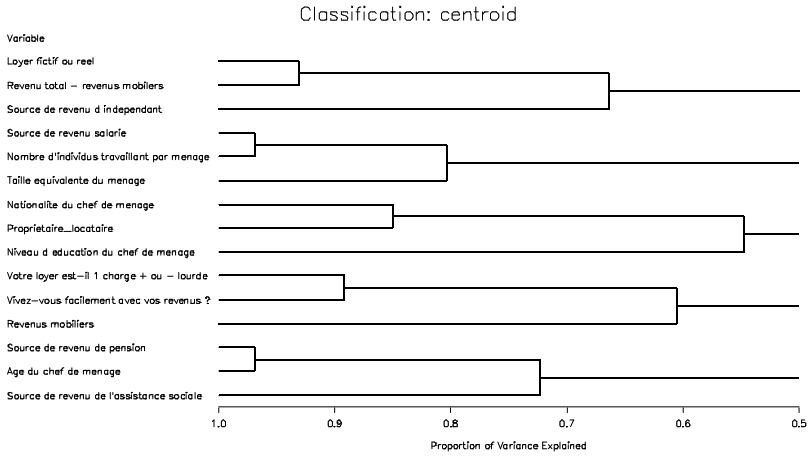
Variable	Revenu mobilier	S_S	S_I	S_P	S_A	Taille Equi.	Nbre Trav.	Age _CM	Niveau. _Educ _CM	Nation. _CM	HS140*	HS120**	prop _loc	loyer génér	Revenu total
Revenu mobilier	1,00	-0,11	0,01	0,07	0,00	-0,01	-0,06	**0,14**	-0,02	-0,03	**0,14**	**0,25**	-0,02	**0,17**	**0,38**
Source _Salaire	-0,11	1,00	0,00	-0,69	-0,07	0,35	0,76	-0,71	0,26	0,25	-0,10	-0,04	0,10	0,28	0,21
Source _Indep.	0,01	0,00	1,00	-0,10	-0,04	0,08	0,12	-0,05	0,09	-0,05	0,03	0,02	-0,04	0,12	0,27
Source _Pension	0,07	-0,69	-0,10	1,00	0,05	-0,17	-0,65	0,77	-0,31	-0,29	0,09	0,01	-0,18	-0,25	-0,21
Source _Assist_ Sociale	0,00	-0,07	-0,04	0,05	1,00	0,04	-0,08	0,02	-0,07	0,01	-0,03	-0,20	0,03	-0,06	-0,12
Taille_ Equival.	-0,01	0,35	0,08	-0,17	0,04	1,00	0,49	-0,13	-0,02	0,08	-0,11	-0,07	-0,17	0,27	0,25
Nbre_de _Trav.	-0,06	0,76	0,12	-0,65	-0,08	0,49	1,00	-0,62	0,25	0,22	-0,07	0,00	0,06	0,30	0,29
Age_chef_ Menage	**0,14**	-0,71	-0,05	0,77	0,02	-0,13	-0,62	1,00	-0,37	-0,28	0,16	0,12	-0,24	-0,20	-0,04
Niv_Educ chef Menage	-0,02	0,26	0,09	-0,31	-0,07	-0,02	0,25	-0,37	1,00	0,22	0,03	0,09	0,16	0,08	0,17
Natio. chef Menage	-0,03	0,25	-0,05	-0,29	0,01	0,08	0,22	-0,28	0,22	1,00	-0,15	-0,09	0,38	0,17	0,05
HS140*	**0,14**	-0,10	0,03	0,09	-0,03	-0,11	-0,07	0,16	0,03	-0,15	1,00	0,41	0,06	-0,07	0,17
HS120**	**0,25**	-0,04	0,02	0,01	-0,20	-0,07	0,00	0,12	0,09	-0,09	0,41	1,00	-0,01	0,15	**0,34**
Proprie_ Loca	-0,02	0,10	-0,04	-0,18	0,03	-0,17	0,06	-0,24	0,16	0,38	0,06	-0,01	1,00	-0,07	-0,08
loyer générique	**0,17**	0,28	0,12	-0,25	-0,06	0,27	0,30	-0,20	0,08	0,17	-0,07	0,15	-0,07	1,00	**0,44**
Revenu total	**0,38**	0,21	0,27	-0,21	-0,12	0,25	0,29	-0,04	0,17	0,05	0,17	**0,34**	-0,08	**0,44**	1,00

*=" *Vos coûts du logements sont-ils une charge lourde"*

**="*Vivez-vous facilement avec vos revenus ?"*

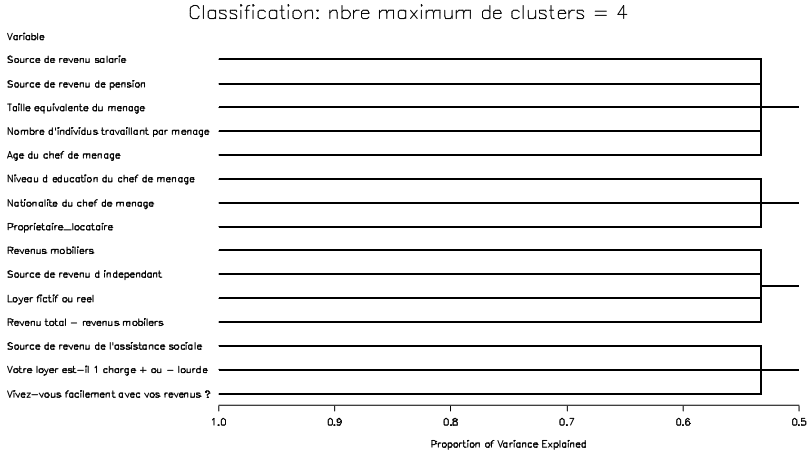
Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Graphique 1 : Classification automatique des variables candidates au statut de variables explicatives et des revenus mobiliers : version centroïd



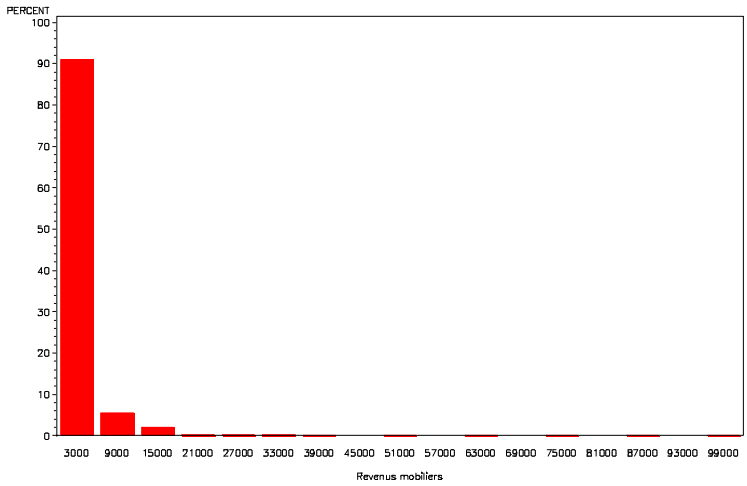
Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Graphique 2 : Classification automatique des variables candidates au statut de variables explicatives et des revenus mobiliers : version nombre maximum de groupes = 4



Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Graphique 3 : Distribution des revenus mobiliers collectés



Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

3. Les quatre procédures d'imputation retenues

Dans ce paragraphe, la présentation des quatre méthodes d'estimation sera suivie par une partie portant sur l'attitude à adopter vis-à-vis des valeurs extrêmes et leur impact éventuel pour chaque méthode. Les quatre méthodes d'estimation qui seront appliquées aux données sont les suivantes : le hot deck aléatoire intra-classes, la médiane intra-classes, le mode intra-classes et la moyenne intra-classes. Elles partagent ensemble des caractéristiques communes qui vont être exposées en premier lieu avant ensuite d'énoncer les principes, puis de décrire la mise en œuvre de chacune des quatre méthodes. Leur choix se justifie par le fait qu'elles se prêtent plus dans des configurations de pauvreté de l'information marquée par la présence d'une faible quantité de variables explicatives candidates. En outre au vu de la nature atypique de la distribution des revenus mobiliers dans le graphique 3 ci-dessus, l'influence des valeurs extrêmes rend problématique le recours aux méthodes de « régression vers la moyenne » c'est-à-dire les modèles linéaires en général. C'est d'ailleurs les limites d'une méthode de régression qui ont amené à chercher une alternative parmi les quatre méthodes précitées. Cette méthode de régression[9] a tendance à imputer les données manquantes par un minimum ou un maximum local en cas de déficit de variance résiduelle.

Les quatre méthodes d'imputation en question partagent un trait commun consistant à imputer les données manquantes par des valeurs de la variable dépendante, c'est à dire les revenus mobiliers, trouvées à l'intérieur des classes auxquelles appartiennent les observations avec des données manquantes. Ces valeurs peuvent être multiples comme dans le cadre du hot deck ou bien sont au plus égales à un pour les trois autres méthodes où les valeurs candidates pour l'imputation sont la médiane, le mode ou la moyenne des observations renseignées des revenus mobiliers de la classe concernée. Quant aux classes d'imputation, elles sont le résultat du croisement des variables explicatives du modèle.

Une fois décrite tout le socle commun aux quatre méthodes d'estimation, il est maintenant possible de définir l'opérateur générique nécessaire à la formulation de la conception de ces méthodes et à la formulation de leur mise œuvre.

Soit X_{ijkl}, une observation i de la classe d'imputation C_{jkl} de n_i observations avec :

- j = 1…5, les 5 quintiles du revenu total ;
- k= 1….4, les classes d'âge du chef de ménage *(16_34 ans, 35_49 ans, 50_64 ans, 65 ans ou +) ;*
- l= 1….3, les 3 modalités de la variable HS140 *("Vos coûts du logement sont-ils charge lourde" :1="Une charge importante", 2="Une charge moyennement importante", 3="Une charge pas du tout importante").*

[9] Il s'agit de la régression généralisée séquentielle du logiciel d'imputation multiple IVEware de l'Université de Michigan.

Pour ne pas alourdir les formulations, aucune mention n'est faite des poids qui ont été effectivement appliqués chaque fois que cela possible dans le processus. Il en sera ainsi sur le reste du document où la somme des X_{ijkl} dans la classe devrait être multipliée par leurs poids respectifs W_i, de même l'effectif n_i d'une classe de répondants devrait être remplacée par l'effectif de la population correspondante estimée à $\sum_{i=1}^{ni} Wi$.

3.1. La méthode du hot deck aléatoire intra-classes

3.1.1. Conception

Il existe plusieurs variantes de la méthode du hot deck [1] [2] [3], la méthode du Plus Proche Voisin (PPV), le hot deck hiérarchisé, le hot deck séquentiel, le hot deck aléatoire intra-classe etc.

La méthode du PPV est une méthode non paramétrique basée sur la distance exprimée en fonction d'autres variables auxiliaires entre un individu non-répondant et un individu répondant. Le donneur retenu pour remplacer la valeur manquante du non-répondant est le répondant le plus proche du non-

répondant $y_i = y_j$ pour certains $j \in s_r$ tel que distance (y_i, y_j) soit minimale, s_r étant l'échantillon des

répondants pour la variable à imputer.

Le hot deck hiérarchisé consiste à choisir les donneurs à partir de la similarité du plus grand nombre possible de variables entre le donneur et le receveur répondant selon un ordre bien déterminé. On élimine au fur et à mesure les variables pour lesquelles les conditions d'égalité ou de proximité ne peuvent être satisfaites

Pour le hot deck séquentiel, le donneur est le répondant précédant le non-répondant dans le fichier trié selon un critère ou un ensemble de critères bien déterminé.

Quant au hot deck aléatoire interclasse, elle consiste à tirer au hasard avec ou sans remise parmi une classe d'observations un donneur pour imputer la donnée manquante. C'est une méthode stochastique. Son avantage en général est de proposer des solutions plausibles et de préserver certains paramètres de la distribution de la variable imputée comme la variance qui est sous-estimée en général dans les méthodes d'estimation déterministes. Par contre, elle ne possède pas les propriétés du maximum de vraisemblance [4]. C'est cette variante qui sera choisie comme modèle de hot deck à appliquer.

3.1.2. Mise en œuvre

Elle consiste à prendre aléatoirement un répondant X_{hjkl} avec remise ou sans remise sur l'ensemble des répondants s_h dans la classe C_{jkl} contenant n_h répondants :

$X_{hjkl} = X_{ijkl}$, $h \in s_h$, tel que $P(X_{ijkl} = X_{hjkl}\ j) = 1/ n_h$

Pour des raisons de commodités, le type de tirage pratiqué est le tirage avec remise car il est non seulement plus facile à programmer, mais il permet aussi d'éviter le risque de se retrouver sans donneur dans une classe à une étape du processus d'imputation. Pour tenir compte de la pondération dans le tirage, chaque donneur est répliqué autant de fois que la partie entière de son poids [5].

3.3. La méthode de la médiane intra-classes

3.3.1 Conception

Grâce à la médiane, cette méthode a la particularité avantageuse d'être sensible aux valeurs extrêmes. Cependant, elle souffre d'une faiblesse qu'elle partage avec les méthodes du mode et de la moyenne : la sous-estimation de la variance car toutes les observations manquantes d'une classe ont la même valeur imputée. La méthode de la médiane est déterministe car c'est la médiane de la classe d'imputation qui est imputée à la donnée manquante y appartenant.

3.3.2 Mise en œuvre

Soit X_{hjkl} une observation de la classe C_{jkl} de répondants d'effectif n_h,
Si n_h est impair, la médiane retenue comme donneur dans cette classe d'imputation est égale à :
$X_{jkl(n_h+1)/2)}$
Si n_h est pair, la médiane retenue comme donneur dans cette classe d'imputation est égale à :
$(X_{jkl(n_h/2)} + X_{jkl[(n_h/2)+1])})/2$

3.4. La méthode du mode intra-classes

3.4.1 Conception

Comme pour la méthode de la médiane intra-classe, la méthode du mode intra-classe est une méthode déterministe. Cependant contrairement à la méthode de la médiane intra-classe, il n'existe pas toujours de solution unique. En l'absence de solution unique, la solution peut être soit multiple, soit inexistante.

3.4.2 Mise en œuvre

Elle consiste à choisir la valeur renseignée la plus fréquente dans la classe d'imputation pour imputer une donnée manquante y appartenant. En cas de solution multiple, c'est la solution par défaut du logiciel SAS qui est adoptée. La médiane de la classe correspondante est utilisée pour pallier l'absence de mode.

3.5. La méthode de la moyenne intra-classes

3.5.1 Conception

Conceptuellement, la moyenne intra-classe est un cas particulier de la régression linéaire généralisée.

Soit y la variable d'intérêt à estimer à partir des variables x_k, k=1…K tel que

$$y_i = a_0 + \sum_{k=1}^{K} ak * xik$$

- Si $a_0 \neq 0$ et $a_k = 0$ pour tout k >= 1, alors le modèle de régression est celui de l'imputation par la moyenne, $y_i = a_0$.

3.5.2 Mise en œuvre

Soit la classe d'imputation C_{jkl} de n_i observations X_{ijkl} avec :

- alors la moyenne[10] dans cette classe est égale à $a_{0i}=(\Sigma_1(i=0)^{\uparrow}ni \square Xijkl)\square$ / n_i

3.6. Choix par rapport aux valeurs extrêmes

La distribution des revenus mobiliers est atypique avec d'une part à l'extrémité inférieure une queue distributions de grande amplitude avec des valeurs très petites composées en général de livrets d'épargne populaire, et d'autre part à l'extrémité supérieure un petit nombre de valeurs très grandes. Cette configuration asymétrique serait de nature à engendrer des biais dans les estimations produites par la méthode de la moyenne intra-classe et par la méthode du hot deck aléatoire intra-classe. Ce risque est nul pour la méthode de la médiane intra-classe et la méthode du mode intra-classe car la médiane et le mode sont insensibles aux valeurs extrêmes. Pour les deux autres méthodes, sensibles aux valeurs extrêmes, déjà évoquées en premier lieu, vu le caractère clivant ou polarisant des revenus mobiliers il a été décidé de n'opérer aucun traitement visant à censurer les valeurs extrêmes. Pour la bonne et simple raison, comme le montre le graphique 3 de la distribution non pondérée ci-dessus, qu'une censure des valeurs extrêmes inférieures entraînerait une «grande» surestimation des revenus mobiliers au voisinage tandis qu'une opération identique pour les valeurs extrêmes supérieures provoquerait une grande sous-estimation des revenus mobiliers au voisinage. En effet dans cette distribution pour des revenus mobiliers allant de 1 à 100 000 Euros, que cela soit pondéré ou non pondéré, le premier quartile est plafonné à 200 Euros et la médiane ne dépasse pas 500 Euros. Dans la queue supérieure de la distribution pondérée ou non-pondérée, le percentile-95 et le percentile-99 sont respectivement inférieurs à 9 000 Euros et à 25 000 Euros.

4. Validation des méthodes d'estimation

La validation des différentes méthodes d'estimation s'opérera à l'aide de trois instruments :

- 1) Le calcul d'un écart moyen absolu (EMA)[11] des résidus qui sont les différences entre les valeurs originales des observations renseignées et les valeurs estimées de ces mêmes observations correspondantes rendues manquantes grâce à la méthode de l'échantillon-test.
- 2) Des tests de cohérence relatifs à la distribution des valeurs mobilières imputées ou non imputées en fonction de certaines variables
- 3) Certains paramètres de la distribution de revenus mobiliers dans leurs composantes selon les flags d'imputation.

4.1. Calcul de l'EMA des résidus avec l'échantillon-test

La méthode de l'échantillon-test consiste à prélever une proportion des observations renseignées d'une variable d'abord pour les rendre manquantes, ensuite pour les imputer et enfin pouvoir calculer des indicateurs sur des approximations des erreurs de prédiction. La particularité de l'échantillon-test adoptée ici est que la proportion est étendue à la totalité, c'est à dire à toutes les observations dont les valeurs mobilières sont renseignées. A classes identiques, si les donneurs pour l'imputation des "données originales" seront les mêmes que durant la phase d'imputation pour les trois autres méthodes, il en sera autrement pour la méthode du hot deck pour lequel on fera appel à un " hot deck inversé". Par " hot deck inversé", il faudrait comprendre la procédure où toutes les valeurs imputées deviennent donneuses pour imputer les valeurs originales transformées en données manquantes pour les besoins de ce test. Il s'ensuivra un calcul de l'EMA pour chaque méthode.

[10] C'est la moyenne pondérée qui est effectivement utilisée. Les poids W_i ne sont pas intégrés dans la formule par souci d'éviter une lourdeur.

[11] En fait, l'EAM permet une mesure approchée du biais et de la variance c'est-à-dire de l'erreur quadratique moyenne.

Tableau 4 : Moyenne pondérée des résidus et Écarts absolus moyens pondérées des résidus des estimations

Indicateurs / Méthodes	Moyenne pondérée des écarts d'erreurs	Ecarts absolus Moyens pondérés
Hot deck	-1135,36	3657,71
Moyenne	-325,30	1912,02
Mode	1045,99	1946,88
Médiane	1049,58	1918,48

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

D'après le tableau 4 ci-dessus, la méthode du hot deck et la méthode de la moyenne disposent de moyennes des écarts négatives, c'est le signe d'une légère surestimation traduisant la supériorité globale des valeurs prédites sur les valeurs observées. Les valeurs des quatre moyennes sont très proches en valeurs absolues à l'exception de celle de la méthode de la moyenne équivalente à -325.30 se particularisant par un plus grand pouvoir de lissage. Si on considère le critère des écarts absolus où ne sont pas prises en compte que les normes des variations indépendamment de leurs signes, les valeurs ne sont pas très éloignées, sauf pour la méthode de hot deck où l'écart est très nettement supérieur à celui des autres méthodes. C'est un résultat prévisible parce que ces méthodes ont tendance à lisser vers le bas et/ou vers le haut les valeurs estimées. Donc, les valeurs extrêmes éligibles pour le hot deck sont susceptibles d'augmenter sensiblement la grande variance.

4.2. Tests de cohérence sur la distribution des revenus mobiliers en fonction d'autres variables

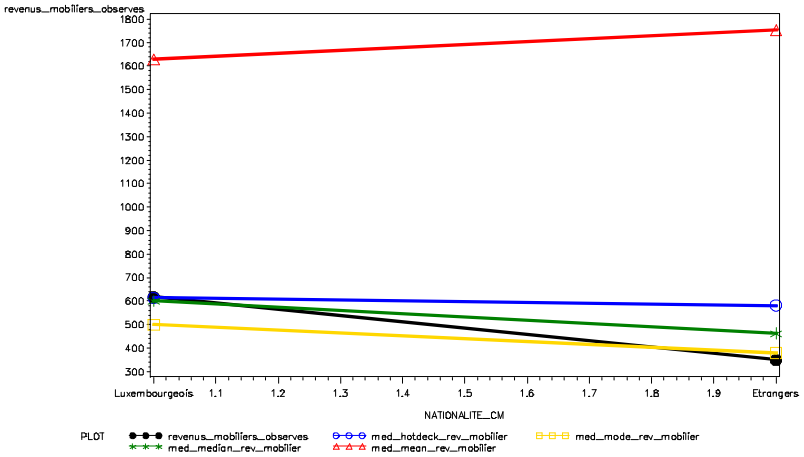
Les variables sollicitées pour tester la cohérence ou la consistance des résultats des quatre méthodes d'estimation sont soit des variables corrélées avec les revenus mobiliers non retenues comme explicatives, soit des variables potentiellement dotées d'un pouvoir discriminant non négligeable par rapport aux revenus mobiliers. L'indicateur de comparaison choisi est la médiane pour éviter l'influence des valeurs extrêmes. L'idée sous-jacente consiste, sous l'adoption implicite de l'hypothèse "Missing At Random" (MAR)[12]sur les non-réponses partielles des revenus mobiliers, de jauger le degré de proximité des profils d'évolution en fonction des modalités des variables-test entre d'une part les données originales, et d'autre part les données imputées à partir des quatre méthodes d'estimation.

Dans la suite, seront énoncées quatre séries de deux graphiques dont un graphique relatif aux revenus mobiliers collectés et l'autre ayant trait aux revenus mobiliers issus des quatre méthodes d'imputation. Les quatre séries portent sur les comparaisons des distributions des revenus mobiliers en fonction des variables que voici dans l'ordre : la nationalité du chef de ménage, l'aisance de vie du ménage avec son niveau de revenu, le statut d'occupation du ménage, le loyer générique.

[12] Un mécanisme de non-réponse est dit "MAR" si le phénomène qui peut être appréhendé et correctement traité à l'aide des variables auxiliaires disponibles.

4.2.1. Distributions des revenus mobiliers en fonction de la nationalité du chef de ménage

Graphique 4 : Évolution de la médiane des revenus mobiliers collectés en fonction de la nationalité du chef de ménage

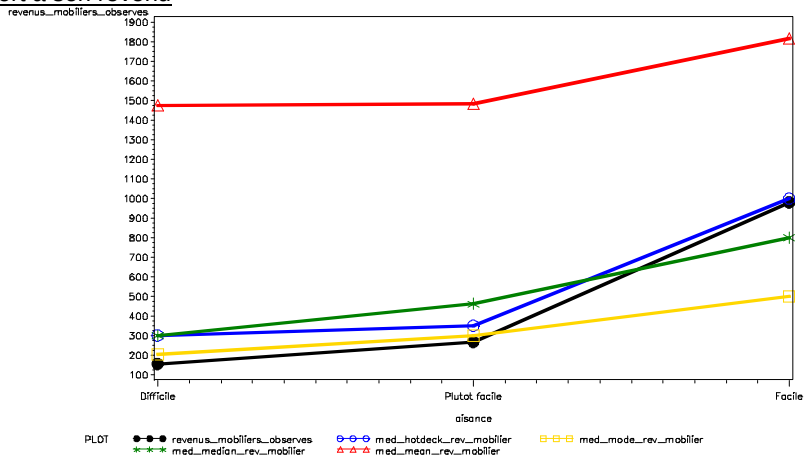


Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Dans le graphique 4 ci-dessus, en se référant à la médiane les revenus mobiliers collectés auprès des Luxembourgeois sont supérieurs à ceux des résidents étrangers. Pour les revenus mobiliers estimés, la méthode du mode et la méthode de la médiane reflètent clairement cette tendance alors que l'évolution reste constante pour la méthode du hot deck. Quant à la méthode de la moyenne, elle ne suit pas la tendance en plus d'être d'un niveau surélevé par rapport aux autres.

4.2.2. Distributions des revenus mobiliers en fonction de l'aisance de vie du ménage avec son niveau de revenu

Graphique 5 : Évolution de la médiane des revenus mobiliers collectés en fonction de l'aisance de vie du ménage par rapport à son revenu

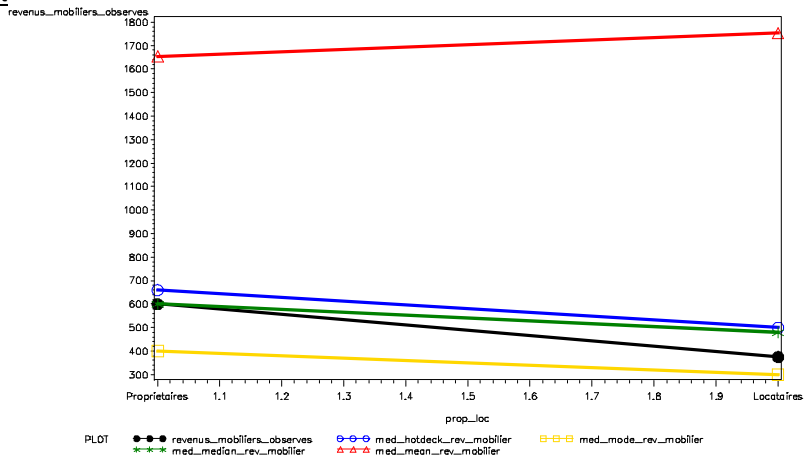


Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Il est logique de supposer que le niveau des revenus mobiliers augmente avec l'aisance avec laquelle vit un ménage comme le graphique 5 ci-dessus semble le corroborer pour les revenus originaux. Il en est de même pour les méthodes imputées dans ce graphique avec une allure très comparable pour la méthode du hot deck, ce phénomène étant moins perceptible pour les autres méthodes encore avec un niveau très élevé pour la méthode de la moyenne.

4.2.3. Distribution des revenus mobiliers en fonction du statut d'occupation du ménage

Graphique 6 : Évolution de la médiane des revenus mobiliers collectés en fonction du statut d'occupation du logement

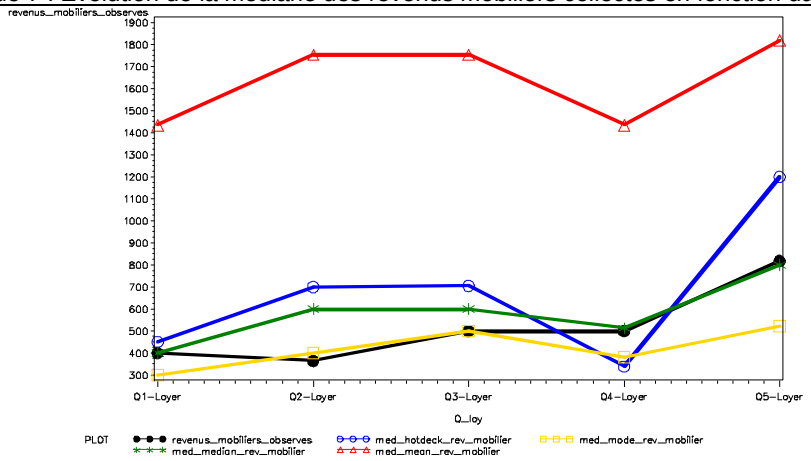


Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

D'après le critère de la médiane, les revenus mobiliers observés des propriétaires sont plus élevés que ceux des locataires comme l'indique le graphique 6 ci-dessus. Selon ce même graphique, seules la méthode du hot deck et la méthode du mode produisent des différences appréciables dans le même sens que les données originales. Par contre, la tendance s'inverse pour la méthode de la moyenne en sus d'un palier encore supérieur à celui des autres méthodes.

4.2.4. Distribution des revenus mobiliers en fonction du loyer générique

Graphique 7 : Évolution de la médiane des revenus mobiliers collectés en fonction du niveau de loyer générique



Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Dans le graphique 10 ci-dessus, les revenus mobiliers collectés croissent avec le quintile des loyers génériques tandis que ce constat n'est pas avéré pour les revenus mobiliers estimés, quelle que soit la méthode d'estimation utilisée.

Sur les quatre tests parcourus ci-dessus :

- En fonction de la nationalité, toutes les méthodes d'estimation ont obéi à l'hypothèse de MAR sauf la méthode du hot deck pour laquelle il y aurait une invariance par rapport à la nationalité.
- En fonction de l'aisance de vie avec le niveau de revenu : les revenus mobiliers imputés avec les quatre méthodes varient dans le même sens que les revenus observés.
- En fonction du statut d'occupation du logement : Seules les courbes de revenus mobiliers estimés avec les méthodes du hot deck et du mode suivent nettement l'allure de la courbe des revenus mobiliers collectés.
- En fonction des quintiles du loyer générique : l'hypothèse MAR n'est vérifiée pour aucune méthode d'estimation des revenus mobiliers.

En fin de compte sur les quatre méthodes, ce sont la méthode du hot deck et la méthode du mode qui ont satisfait dans la plupart des cas aux critères de l'hypothèse du MAR.

4.3. Analyse de quelques paramètres de la distribution des revenus mobiliers

Dans cette dernière phase de validation, les statistiques suivantes pour les données collectées et pour les données estimées à partir des quatre méthodes seront calculées et comparées : le minimum, le maximum, la moyenne, la médiane et l'écart-type dans le tableau 5 ci-dessous.

Concernant les minima et le maxima dans ce tableau, l'étendue des revenus mobiliers imputés avec la méthode du hot deck épouse le mieux le profil de l'étendue des revenus mobiliers collectés. Cela est d'autant plus compréhensible que les autres méthodes d'estimation « lissent » les valeurs extrêmes.

Quant aux moyennes issues des revenus mobiliers imputés dans le même tableau 5 ci-dessous, elles sont toutes très proches à celle des revenus mobiliers observés à l'exception de celle calculée avec les valeurs imputées par la méthode de la moyenne et par la méthode du hot deck. La supériorité de la moyenne pour la méthode du hot deck serait causée l'élection de donneurs parmi les valeurs extrêmes supérieures qui sont de fait censurées par construction dans les autres méthodes.

Pour les médianes, il n'y a pas de grandes différences notables entre elles à part la médiane des revenus mobiliers estimés à partir de la méthode de la moyenne.

Enfin pour les écart-types, celui des revenus mobiliers collectés est très inférieur à celui issu des estimations par la méthode du hot deck. Par contre, il est très supérieur à ceux issus de revenus mobiliers estimés à l'aide des autres méthodes. Ces différences étayent une sous-estimation vraisemblable de la variance par ces trois autres méthodes d'estimation à l'opposé de la méthode du hot deck.

En résumé parmi les cinq paramètres étudiés du tableau 5 ci-dessous, sur une échelle de qualité de performance dans un sens croissant, la méthode de la moyenne et la méthode du hot deck constituent respectivement la borne inférieure et la borne supérieure. La méthode du mode et la méthode de la médiane gravitent entre ces bornes.

Tableau 5 : Statistiques sur la distribution des données originales et sur les données estimées avec les quatre méthodes d'estimation.

Indicateurs / Méthodes	Minimum	Maximum	Moyenne	Médiane	Ecart-type
Valeurs observées	1	100000	1904	500	32894
Hot deck	5	72000	2989	606	42684
Moyenne	33	53845	2811	1662	26979
Mode	20	53845	1979	400	27953
Médiane	33	53845	1873	566	27612

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

5. Conclusion

Pour la validation des quatre méthodes d'estimation des revenus mobiliers, trois types de tests ont été sollicités : un test sur les résidus des erreurs pour les différentes estimations, un test sur la plausibilité de la distribution des différents revenus mobiliers imputés en fonction de certaines variables et un test de comparaison des paramètres respectifs des distributions des revenus mobiliers observés et des distributions des revenus imputés.

Le premier test sur les résidus consacre la supériorité des performances de la méthode du hot deck. Dans le second test sur la plausibilité selon l'hypothèse MAR des distributions des revenus mobiliers, la méthode du hot deck et la méthode du mode donnent des résultats plus cohérents. Enfin dans le dernier test, l'avantage revient à la méthode du hot deck, les performances de la méthode de la moyenne sont jugées largement inférieures à celles des autres méthodes.

A l'aune de tous ces tests, il en ressort que c'est la méthode du hot deck qui paraît la moins mauvaise. L'impact de cette méthode d'imputation serait in fine une relative surestimation des revenus mobiliers en se fiant au critère de variation de la médiane des valeurs imputées. C'est le paramètre le plus adapté pour juger de la qualité de l'estimation d'une variable dont la distribution est fortement affectée par les valeurs extrêmes. Cette surestimation serait plus importante pour l'estimateur Horvitz-Thompson du total des revenus mobiliers bruts sur la base de la différence significative entre la moyenne des revenus mobiliers observés et celle des revenus mobiliers estimés avec la méthode du hot deck.

Bibliographie

[1] CARON N., Les principales techniques de correction de la non-réponse et les modèles associés, Série des Documents de Travail « Méthodologie Statistique » INSEE N°9604, janvier 1996, Paris.
[2] HAZIZA D., Inférence en présence d'imputation : un survol, Journées de Méthodologie Statistique, 16-17 décembre 2002 à Paris.
[3] Fuller W.A., Kim J.K., Imputation hot deck pour le modèle de réponse, Techniques d'enquête Vol. 31, No 2, pp. 153-164 Statistique Canada, No 12-001 au catalogue, Décembre 2005.
[4] Schulte Nordholt E., Imputation: Methods, Simulation Experiments and Pratical Examples International Statistical Review / Revue Internationale de Statistique, Vol. 66, No. 2 (Aug., 1998), pp. 157-180.
[5] Clayton D., Mander A., Weighted Hotdeck Imputation, http://fmwww.bc.edu/repec/bocode/w/whotdeck.pdf.

Annexe :

Liste des variables explicatives candidates pour les modèles et la variable dépendante

Age_cm="Age du chef de ménage"
Niveau_educ_cm="Niveau d'éducation du chef de ménage"
Loyer_gener="Loyer fictif ou réel"
revenu_global="Revenu total moins la composante mobilière au niveau du ménage"
HS120="Vivez-vous facilement avec vos revenus ?"
HS140=" Vos coûts du logement sont-ils une charge lourde"
nationalite_cm="Nationalité du chef de ménage"
Prop_loc="Propriétaire ou Locataire"
HPY052="Revenus mobiliers"
Sourcea="Source de revenu d'assistance sociale"
Sourcei="Source de revenu d'indépendant"
Sourcep="Source de revenu de pension"
Sources="Source de revenu salarié"
Taill_eq="Taille en équivalents du ménage"
Nb_trav="Nombre d'individus travaillant par ménage"

Modalités de la variable HS120 (*"Vivez-vous facilement avec vos revenus ?"*)

1="Très difficilement"
2="Plutôt difficilement"
3="Difficilement"
4="Plutôt facilement"
5="Facilement"
6="Très facilement";

Modalités de la variable HS140 (*"Vos coûts du logement sont-ils une charge lourde"*)

1="Une charge importante"
2="Une charge moyennement importante"
3="Une charge pas du tout importante";

Distributions des observations donneuses et des observations receveuses ainsi que leurs ratios

Cellule d'imputation	Fréquence Receveuse non pondérée	Fréquence Donneuse non pondérée	Fréquence Receveuse pondérée	Fréquence Donneuse pondérée	Ratio Receveur/ Donneur non pondéré	Ratio Receveur/ Donneur pondéré
1	13	25	247	530	0,52	0,47
2	15	27	791	1080	0,56	0,73
3	5	12	259	756	0,42	0,34
4	9	10	602	841	0,90	0,72
5	8	21	282	725	0,38	0,39
6	12	26	528	1088	0,46	0,49
7	5	11	153	689	0,45	0,22
8	11	7	1215	416	1,57	2,92
9	5	8	101	489	0,63	0,21
10	6	7	348	189	0,86	1,84
11	3	6	185	347	0,50	0,53
12	1	9	173	1297	0,11	0,13
13	8	30	545	948	0,27	0,57
14	19	20	710	1194	0,95	0,59
15	12	14	418	684	0,86	0,61
16	18	10	1537	775	1,80	1,98
17	14	32	628	1258	0,44	0,50
18	8	27	627	1905	0,30	0,33
19	17	17	1147	1103	1,00	1,04
20	18	27	1837	2806	0,67	0,65
21	3	4	122	84	0,75	1,46
22	3	2	119	15	1,50	8,07
23	4	2	258	162	2,00	1,60
24	7	18	727	1681	0,39	0,43
25	5	19	315	842	0,26	0,37
26	9	19	750	1572	0,47	0,48
27	21	13	974	898	1,62	1,09
28	10	13	335	961	0,77	0,35
29	10	37	545	1536	0,27	0,36
30	19	38	967	2521	0,50	0,38
31	20	35	973	1611	0,57	0,60
32	27	36	1839	2542	0,75	0,72
33	7	15	280	506	0,47	0,55
34	6	5	827	288	1,20	2,88
35	8	11	621	598	0,73	1,04
36	12	14	568	852	0,86	0,67
37	6	13	166	488	0,46	0,34
38	18	16	1016	1163	1,13	0,87
39	13	8	519	165	1,63	3,15
40	6	4	377	210	1,50	1,79
41	8	25	295	1108	0,32	0,27
42	34	46	2690	3907	0,74	0,69
43	24	37	1415	2431	0,65	0,58
44	10	26	559	1365	0,38	0,41
45	4	16	97	705	0,25	0,14

	46	11	20	480	2177	0,55	0,22
	47	18	19	1010	1439	0,95	0,70
	48	12	22	437	1109	0,55	0,39
	49	8	11	193	489	0,73	0,40
	50	14	22	525	1413	0,64	0,37
	51	10	12	628	638	0,83	0,98
	52	3	6	12	326	0,50	0,04
	53	14	25	369	1191	0,56	0,31
	54	37	81	1723	3830	0,46	0,45
	55	43	48	2368	2430	0,90	0,97
	56	12	13	479	1010	0,92	0,47
	57	5	11	35	644	0,45	0,05
	58	15	37	750	1660	0,41	0,45
	59	31	56	1809	3856	0,55	0,47
	60	6	13	411	909	0,46	0,45
Total		740	1214	40917	70449		

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC