Comparaison de quatre méthodes d'imputation des revenus mobiliers dans le cadre de l'enquête EU-SILC[1]

Modou DIA[2]

Problématique

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. Cependant, les niveaux de difficulté de la tâche dépendent de la nature de la composante du revenu avec son profil singulier. 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 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 précise le montant des allocations perçues en fonction de la composition du ménage. Par contre, tout autre est l'ampleur des difficultés dès qu'il s'agit de traiter de l'imputation des revenus mobiliers qui sont l'objet de cette communication. Ces difficultés sont tributaires de certaines caractéristiques propres à cette composante de revenu.

Caractéristiques des revenus mobiliers et principaux défis pour leur imputation

D'abord, les revenus mobiliers sont composés d'un portefeuille de produits très divers avec une complexité et une volatilité très variables. Ensuite, leur taux de données manquantes est relativement élevé. 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. D'où l'importance à attacher au choix des modèles d'imputation et à leurs modes de validation.

Méthodes et Validation

Les quatre méthodes d'imputation qui sont retenues dans cette présentation sont les suivantes : le hotdeck aléatoire intra-classes, la médiane intra-classes, le mode intra-classes et la moyenne intra-classes. A partir de ces méthodes classiques, l'originalité réside dans les innovations introduites tant au niveau de leur mise en œuvre qu'au niveau de leur validation. Dans la mise en œuvre, outre le recours à certaines variables subjectives pertinentes, seront expérimentées dans les modèles d'autres variables découlant respectivement d'une approche pseudo-récursive ou d'un calcul d'un revenu "fictif" comme respectivement les quantiles du revenu total du ménage (sans la composante mobilière) ou les quantiles du loyer imputé du ménage.

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

[2]CEPS/INSTEAD, 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/**

Quant à la validation, elle reposera sur trois piliers :

- La méthode de l'échantillon-test étendue 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 hotdeck par rapport auquel on fera appel à un "hotdeck inversé" à partir des valeurs estimées des valeurs manquantes. Il s'ensuivra un calcul de l'erreur quadratique moyenne pour chaque méthode.
- La distribution des revenus mobiliers (imputés et/ou observés) en fonction de différentes variables potentiellement discriminantes comme le statut d'occupation par exemple.
- La variance des revenus mobiliers (imputés et/ou observés) pour chaque méthode.

Quant aux sources de données pour l'application de ces méthodes, elles proviennent de la sixième vague de l'EU-SILC pour le Luxembourg relative à l'année 2008 qui comporte 3779 observations de ménages privés au sein desquelles des informations sur les revenus mobiliers ont été recueillies.

Conclusion

Elle consistera à faire la synthèse des résultats des trois tests ci-dessus pour déterminer la moins mauvaise méthode sachant que le critère de la variance la plus faible ne sera pas forcément décisif dans la mesure où une méthode, comme celle de la moyenne intra-classe, a tendance à sous-estimer la variance.