RECONSTITUER LA FACTURE D'ÉNERGIE TOTALE DES MÉNAGES À PARTIR DE DEUX SOURCES DISTINCTES (ENL ET ENTD) : LE PSEUDO-APPARIEMENT DANS LE MODÈLE DE MICROSIMULATION PROMETHEUS

Mathilde CLÉMENT, Charles PEROUMAL, Annaïck ROLLAND

Ministère de la Transition Écologique et Solidaire, CGDD, Service de l'économie, de l'évaluation et de l'intégration du développement durable (Seeidd)

mathilde.clement@developpement-durable.gouv.fr
charles.peroumal@developpement-durable.gouv.fr
annaick.rolland@developpement-durable.gouv.fr

Mots-clés : pseudo-appariement (statistical matching), stratification et hot-deck aléatoire

Résumé

Cet article présente la principale étape de constitution du modèle de micro-simulation *Prometheus* du Seeidd : le pseudo-appariement entre les deux enquêtes ménages que sont l'ENL (Enquête nationale logement de l'Insee) et l'ENTD (Enquête nationale des transports et des déplacements du SDES). Le modèle de micro-simulation *Prometheus* vise en effet à estimer les dépenses énergétiques totales des ménages aussi bien pour le poste logement (chauffage, cuisson, électricité) que pour le poste transport (carburants des véhicules), dans l'objectif de simuler l'impact des réformes de fiscalité énergétique sur les factures des ménages. Le pseudo-appariement de l'ENL et l'ENTD permet de disposer d'une base de donnée de l'ordre de 27 000 ménages dans laquelle on peut calculer, pour chaque ménage, les consommations annuelles de combustibles et carburants en fonction des caractéristiques des ménages, de leurs logements et de leurs véhicules .

Plus précisément, le pseudo-appariement de l'ENL et de l'ENTD est réalisé par stratification et hot-deck aléatoire. Il s'agit, à partir de variables pertinentes et communes aux deux enquêtes, d'attribuer à chaque ménage de l'ENL le parc de véhicules motorisés d'un ménage de l'ENTD ayant des caractéristiques proches. Sont également appariées les caractéristiques de ces véhicules en termes de carburant, consommation et de mobilité (kilométrage annuel). La qualité finale de l'appariement repose sur l'hypothèse d'indépendance conditionnelle : elle stipule que toutes les variables de l'ENL qui peuvent influencer les variables de l'ENTD appariées (en termes de mobilité ou de véhicules) sont prises en compte dans la stratification.

Dans un premier temps, on constitue les strates retenues pour l'appariement. Pour ce faire, on reconstruit un maximum de variables et modalités communes entre l'ENL et l'ENTD, et on compare les distributions de ces variables, afin de vérifier qu'elles sont proches dans les deux sources. On détermine ensuite celles qui expliquent le mieux le kilométrage annuel parcouru par l'ensemble des véhicules du ménage (via une régression linéaire), qui est notre principale variable d'intérêt pour reconstituer les dépenses de carburants du ménage dans le modèle. Les 11 régresseurs retenus sont hiérarchisés par ordre décroissant de pouvoir explicatif. Cet ordonnancement sert de base au découpage par strate utilisé pour le tirage d'un donneur. *In fine*, les variables de stratification portent sur le nombre de véhicules du ménage (c'est la seule variable relative à l'équipement des

ménages en véhicules motorisés disponible dans l'enquête Logement), la durée hebdomadaire des trajets domicile-travail de la personne de référence et de son conjoint éventuel s'ils sont réalisés habituellement en voiture, et les principales caractéristiques socio-démographiques du ménage qui influent sur ses déplacements en véhicule motorisé (type d'unité urbaine, age, niveau de revenus, situation d'activité, niveau de diplôme, etc.).

Dans un second temps, on procède à l'appariement : à chaque ménage de l'ENL est attribué via un tirage aléatoire simple avec remise un ménage de l'ENTD appartenant à la même strate. Lorsque la strate des donneurs est vide, ou que son cardinal est inférieur à 5, un donneur est tiré dans la strate plus large constituée par les 10 premières variables de stratification mais excluant la 11^{e}. L'opération est répétée de manière récursive jusqu'à appariement de tous les ménages de l'ENL. *In fine*, 96 % des ménages de l'ENL sont appariés sur des strates constituées par 5 variables ou plus sur les 11 variables de stratification initialement retenues. La condition de taille minimum sur la strate de donneurs permet d'éviter de tirer le même donneur de nombreuses fois et de déformer la distribution des kilométrages annuels.

La dernière phase consiste à vérifier la validité de l'appariement en comparant les caractéristiques de mobilité des ménages et du parc de véhicules avant et après appariement. En particulier la structure de carburants (essence ou diesel) du parc automobile dans la base ENL post appariement reste proche de celle de l'ENTD, en dépit du fait que le nombre de véhicules diesels possédés par les ménages n'est pas une variable de stratification puisque cette information n'est pas disponible dans l'ENL.

Bibliographie

[1] Leulescu Aura, Agafitei Mihaela, « Statistical matching: a model based approach for data integration », *Eurostat Methodologies and Working Papers*, 2013
[2] de Waal Ton, « Statistical matching: Experimental results and future research questions », *CBS Discussion Paper,* 2015.