Choix des variables auxiliaires pour le redressement d'une enquête de mobilité

Fabio Rendina[1] & Mathieu Rabaud[2] & Myriam Maumy-Bertrand[3] & Jimmy Armoogum[4]

[1]*IFSTTAR, AME-DEST,*
14-20 Boulevard Newton, Champs sur Marne, F-77447 Marne la Vallée Cedex 2
Email: fabio.rendina@ifsttar.fr

[2]*CEREMA,*
Direction territoriale Nord-Picardie - 44 ter, rue Jean Bart CS 20275 - 59019 Lille Cedex
Email: Mathieu.Rabaud@cerema.fr

[3]*UNIVERSITÉ DE STRASBOURG,*
7 rue René Descartes, 67084 Strasbourg
Email : mmaumy@math.unistra.fr

[4]*IFSTTAR, AME-DEST,*
14-20 Boulevard Newton, Champs sur Marne, F-77447 Marne la Vallée Cedex 2
Email: jimmy.armoogum@ifsttar.fr

Mots-clés. Sondage, Calage sur marge, Comparaison de calage.

La qualité des estimations issues d'une enquête par sondage peut être améliorée en présence d'information auxiliaire [Yves Tillé (1992)]. Le calage sur marges est une méthode de redressement d'enquête très efficace lorsque la taille de l'échantillon est suffisamment grande [Jean-Claude Deville, Carl-Eric Särndal (1992)]. Sous l'influence d'un trop grand nombre de variables auxiliaires, le redressement peut conduire à des instabilités des poids provoquant ainsi une diminution de la précision des estimations. Le but est le choix judicieux des variables auxiliaires jugées pertinentes afin d'améliorer la qualité des estimations obtenues. La précision des estimations sera quantifiée par le calcul de la variance des estimateurs calés obtenus.

A partir des années 60, les premières enquêtes ménages et déplacements ont vu le jour. Le but des enquêtes ménages déplacements est de prendre connaissance de la mobilité urbaine, essentielle pour la construction et le suivi de politiques urbaines. Les enquêtes sont renouvelées en moyenne de manière décennale (1976, 1983, 1991, 2001 et 2011 pour le cas de l'Ile-de-France). Cette méthodologie est actuellement toujours utilisée, ce qui permet la comparaison des mêmes variables (nombre de voitures privées, nombre de deux roues motorisés, distance parcourue par déplacement, etc.) au fil du temps.

Nous disposons d'une base de données comportant dix-neuf enquêtes ménages et déplacements (EMD) qui ont étés rassemblées dans une même base unifiée. Voici les différentes EMD :
Nice 2009, Marseille 2009, Bordeaux/Gironde 2009, Strasbourg 2009, Grenoble 2010, Saint-Étienne 2010, Bayonne 2010, Saint-Quentin-en-Yvelines 2010, Amiens 2010, Caen 2011, Valenciennes 2011, Île-de-France 2011, Angers 2012, Douai 2012, Clermont-Ferrand 2012, Toulouse 2013, Nancy 2013, Valence 2014, Montpellier 2014.

Cela fut possible car toutes les enquêtes suivent la méthodologie appelée « Standard CERTU ». Cette méthodologie porte le nom de « Standard CERTU » car elle provient du Centre d'Etudes sur les Réseaux les Transports, l'Urbanisme et les constructions publiques (CERTU). Ce centre à ce jour a intégré le Centre d'Etudes et d'expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement (CEREMA) tout comme le SETRA. Le financement des enquêtes provient d'un partenariat associant les collectivités locales (structure intercommunale, région, département) ainsi que l'Etat (qui subventionne environ 20%). Les enquêtes portent sur un échantillon représentatif de ménages correctement dispersés sur l'aire d'étude. La taille de l'échantillon est fixée de manière à assurer une fiabilité des résultats permettant une analyse sectorielle.

Deville et Särndal (1992) ont généralisé les estimations par régression en introduisant les estimations par « calage sur marge ». Ils se sont appuyés sur les travaux d'ajustement de sondage via les données du recensement de Lemel (1976).
Le but du calage sur marge est d'améliorer l'estimation de la variable d'intérêt en réduisant la variance de l'estimateur. La réduction de la variance est liée à la corrélation de la variable auxiliaire avec la variable d'intérêt.
La technique consiste à faire concorder les marges de l'échantillon avec celle de la population en appliquant une nouvelle pondération à l'échantillon.

Les variables suivantes sont connues par expérience sur le comportement de la mobilité des individus, il serait donc intéressant de les considérer pour le calage :

Pour les INDIVIDUS	**Pour les MENAGES**
Sexe	Motorisation
Age	Taille du ménage
Niveau d'étude	Type d'habitat[1]
Statut[2]	Age de la personne de référence

Lors de la sélection des variables plusieurs critères rentrent en compte. Premièrement, les modalités de réponse des variables doivent être communes entre le recensement et la base unifiée. Il est possible de devoir regrouper des modalités dans l'une des bases de données pour retrouver les mêmes modalités de réponse. Par exemple, le recensement nous donne les classes d'âge suivantes

[1] Locataire, locataire HLM ou propriétaire

[2] Personne active ou inactive

(00-05 ans, 06-10 ans, 11-17 ans, 18-24 ans, 25-39 ans, 40-54 ans, 55-64 ans, plus de 65 ans), alors que l'âge précis de chaque individu est donné dans la base unifiée. On introduit donc ces classes d'âges dans la base unifiée non bien difficilement pour obtenir la compatibilité parfaite. Le type d'habitat devient donc une variable non utilisable pour des problèmes de compatibilité.
Deuxièmement, il faut s'assurer de la compatibilité entre les modalités de réponse entre le recensement et la base unifiée. On aperçoit que le niveau d'étude n'est plus une variable utilisable car au fil du temps, les diplômes ont évolué. Par exemple, le baccalauréat n'est pas comparable entre les personnes âgées et les jeunes bacheliers.
Troisièmement, la personne de référence étant clairement identifiée dans le recensement, elle est malheureusement biaisée dans la base unifiée. Les réponses à cette question ne sont donc pas comparables. La variable de l'âge de la personne de référence n'est donc plus utilisable.

Lors de cette étude, nous avons mis en avant différentes variables auxiliaires ayant un impact sur la mobilité des individus. En supposant, que le nombre de déplacements par jour soit bon indicateur de la mobilité, l'étude montre que le calage le plus pertinent est le calage via la Taille des ménages et la motorisation au niveau des ménages, puis l'âge, le sexe et le statut au niveau des individus.

Nous devons toutefois préciser que, nous n'avons considéré qu'une seule variable d'intérêt pour modéliser la mobilité. Il aurait été intéressant d'estimer la variance des estimateurs calés via des variables d'intérêts telles que le nombre de véhicules possédés ou le nombre de déplacement en transport en commun par jour. Nous avons opéré à un regroupement D10 sur nos secteurs de tirage. Pour les plus grosses enquêtes, ce regroupement est trop lourd. Rappelons que le regroupement de secteur de tirage peut dans certains cas regrouper des secteurs totalement différents et donc engendrer une perte importante d'informations. Il serait donc judicieux d'obtenir un découpage D30 pour les plus grandes enquêtes (nombre de secteurs de tirage > 90) et un découpage D20 pour les enquêtes moyennement grandes (nombre de secteurs de tirage entre 60 et 90).
La qualité des estimations issues d'une enquête par sondage peut être améliorée en présence d'information auxiliaire [Tillé Y (1992)]. En disposant d'un grand nombre de variables auxiliaires, le redressement par calage peut conduire à des instabilités des poids provoquant ainsi une diminution de la précision des estimations. Cette communication a pour but, d'une part, de mener une discussion sur le choix des variables auxiliaires à utiliser lors d'un redressement et d'autre part, de proposer une optimisation du redressement par calage d'une enquête sur la mobilité. Disposant de nombreuses variables auxiliaires connues, le travail consistera à éviter le phénomène dit de « sur-calage » : des difficultés à satisfaire les équations de calage et une saturation du nombre de degré de libertés ayant pour conséquences une explosion des poids et une augmentation de la variance, voire une impossibilité de calculer les poids de calage. La variance d'un estimateur est un excellent outil pour quantifier la précision d'un estimateur donné. L'obtention de la variance minimale pour une variable d'intérêt donnée dépend des variables auxiliaires choisies. Les variables auxiliaires minimisant la variance peuvent donc être différentes d'une variable d'intérêt à une autre. En considérant plusieurs variables d'intérêt d'un même sujet, pas forcément très corrélées, il faudra proposer une méthode pour sélectionner les variables auxiliaires qui permettent d'établir un système de pondération unique pour différentes variables d'intérêt d'un même thème.

Bibliographie

[1] Deville, J.-C. et Särndal, C.E. (1994) : Variance estimation for the regression imputed Horvitz-Thompson estimator, Journal of Official Statistics, Vol. 10, N° 4 pp. 381-394.

[2] Deville, J.-C. , Särndal, C.E. et Sautory, O. (1993): Generalised raking procedures in survey sampling, Journal of the Americain Statistical Association, Vol. 88, pp. 1013-1020.

[3] Lemel, Y. (1976) : Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondages, Annales de l'Insee, N°22-23, pp. 273-281.

[4] Tillé, Y. (1992). Utilisation a posteriori d'informations auxiliaires en théorie des sondages sans référence à un modèle. Ph.D. thesis, Université Libre de Bruxelles.