Optimisation du redressement d'un sous-échantillon d'une enquête : Application à un sous-échantillon de l'Enquête Nationale sur les Transports et les Déplacements de 2007-2008

Toky Randrianasolo, Jimmy Armoogum

Optimisation du redressement d'un sous-échantillon d'une enquête : Application à un sous-échantillon de l'Enquête Nationale sur les Transports et les Déplacements de 2007–2008

Optimization of a survey sub-sample reweighting: Application to a sub-sample of the 2007–2008 French National Travel Survey

Toky Randrianasolo[a], Jimmy Armoogum[b]

[a]*La Poste, Branche Service-Courrier-Colis, Direction Financière, Direction de la Comptabilité et des Statistiques, Direction des Statistiques, Département Conception et Pilotage des Études Statistiques, CS 50017, F-93192 Noisy-le-Grand Cedex*
[b]*Université Paris-Est, Institut Français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR), Département Aménagement Mobilités et Environnement (AME), Laboratoire Dynamiques Économiques et Sociales des Transports (DEST), Bâtiment Bienvenüe, Cité Descartes, Champs-sur-Marne, 14–20 boulevard Newton F-77447 Marne-la-Vallée Cedex 2*

Toky Randrianasolo
Méthodologue
E-mail : toky.randrianasolo@laposte.fr
Tél. : +33 (0)1 84 09 33 26.
Auteur correspondant

Jimmy Armoogum
Chargé de recherche
E-mail : jimmy.armoogum@ifsttar.fr
Tél. : +33 (0)1 81 66 86 07.

Résumé

Nous proposons une méthode d'optimisation du redressement d'un sous-échantillon d'une enquête. L'objectif est d'éviter une sur-dispersion des poids de calage et des grandes variances, induites par un petit nombre de degrés de liberté lorsque le nombre de variables auxiliaires est grand. Les variables auxiliaires sont choisies de telle sorte à avoir une précision satisfaisante des estimations au niveau du domaine. Les variances sont estimées à l'aide de la méthode proposée par Deville et Särndal (1992) formalisant le calage sur marges. La méthode est appliquée à un sous-échantillon de l'Enquête Nationale sur les Transports et les Déplacements de 2007-2008.

Mots-clés : enquête transport, sous-échantillon, redressement, variables auxiliaires, poids de calage

Abstract

We propose a method of optimizing the reweighting of a survey sub-sample. The objective is to avoid an over-dispersion of the calibration weights and too large variances, induced by a small number of degrees of freedom when the number of auxiliary variables is large. Auxiliary variables are chosen so as to obtain a satisfactory precision of the estimates at the domain level. Variances are estimated by means of the method proposed by Deville and Särndal (1992) which formalises the calibration techniques. The method is applied to a sub-sample of the 2007-2008 French National Travel Survey.

Keywords: travel survey, sub-sample, reweighting, auxiliary variables, calibration weights

Introduction

La qualité des estimations issues d'une enquête par sondage peut être améliorée en présence d'information auxiliaire (Tillé, 1992). L'information auxiliaire est un regroupement de variables (quantitatives ou qualitatives) disponibles pour toute la population. Ainsi, les recensements et les registres de population sont de grandes sources d'information auxiliaire. Connues au niveau de toute la population, ces variables dites auxiliaires peuvent alors être directement utilisées dans les formules des estimateurs, notamment dans les formules des estimateurs par calage.

Le calage sur marges permet en effet de redresser efficacement une enquête lorsque la taille de l'échantillon est suffisamment grande (Deville et Särndal, 1992). Généralement, nous redressons une enquête par rapport à quelques variables auxiliaires disponibles, sans regarder l'impact de ce redressement sur la précision de l'estimateur. Ce papier propose une optimisation du redressement en tenant compte de la qualité des estimations fournies.

Lorsque le nombre de variables auxiliaires disponibles est grand et que l'on se restreint à un sous-échantillon, le redressement par calage peut conduire à des instabilités des poids, provoquant ainsi une diminution de la précision des estimations (voir, par exemple, Chauvet et Goga, 2012). Cet article a pour but de mener une discussion sur le choix des variables auxiliaires à utiliser lors d'un redressement au niveau d'un sous-échantillon. Dans le cas de l'estimateur par calage, l'obtention d'une précision minimale pour une variable d'intérêt donnée dépend des variables auxiliaires choisies. Les variables auxiliaires minimisant la précision (et donc, 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, nous proposons 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.

Un rappel sur le principe du calage sur marges est donné dans la Section 1. Dans la Section 2, nous donnons une présentation de l'Enquête Nationale sur les Transports et les Déplacements 2007–2008 ainsi que le contexte du redressement du sous-échantillon Rhône-Alpes. La Section 3 présente la méthode proposée afin de sélectionner les variables auxiliaires nécessaires qui maximisent la précision des estimations et ouvre une discussion sur les résultats obtenus.

1. Les différents types de biais d'une enquête

Selon Razafindranovona (2015), nous pouvons classer les types d'erreur en sondage en quatre principales catégories :

- l'erreur due à la base de sondage (lorsque celle-ci ne contient pas tous les éléments de la population cible)
- l'erreur due à l'échantillonnage (le fait de prendre une réalisation d'un tirage aléatoire) ;
- l'erreur due à la mesure (lorsque l'on observe pour un individu et pour une variable d'intérêt, une valeur différente de sa vraie valeur) ;
- l'erreur due à la non-réponse (le fait de l'incapacité de mesurer sur toutes les unités de l'échantillon toutes les variables d'intérêt).

Les erreurs de mesure sont compliquées à détecter et à corriger, car il s'agit généralement des omissions. Soit l'enquêté a oublié de déclarer un ou plusieurs évènements, soit il ne souhaite pas les décrire.

Après la phase de collecte des données, nous pouvons classer les techniques de correction de la non-réponse en deux catégories qui cohabitent dans la pratique (Deville et Särndal, 1992 ; Rao et Singh, 2009 ; Goga et al, 2011) :

- l'imputation : cette technique est généralement utilisée pour amender la non-réponse partielle et les erreurs de mesure ;
- la repondération des répondants : cette technique est surtout utilisée pour corriger les défauts de la base de sondage, les défauts de l'échantillonnage, et la non-réponse totale.

Selon Emrich (1983), la non-réponse totale ajoute une phase supplémentaire dans l'échantillonnage. En effet, on tire d'abord un échantillon dans la population selon un plan sondage connu. Puis on considère que l'ensemble des répondants est issu d'un tirage selon un plan de sondage inconnu conditionnellement à l'échantillon de départ : le mécanisme de réponse (Beaumont, 2005).
Le mécanisme de réponse peut dépendre de variables qui sont disponibles (qui existent, par exemple, dans le recensement), de variables qui ne sont pas disponibles (par exemple, si le logement dispose d'un interphone ou non), mais aussi des variables d'intérêt. Si le mécanisme de réponse dépend surtout des variables dont on ne dispose pas ou bien des variables d'intérêt, alors les estimations seront fortement biaisées. Le redressement pourra éventuellement diminuer ces biais, dès lors que les variables du mécanisme de réponse sont corrélées avec les variables dont on dispose (Särndal et Swensson, 1987). Tout l'art du redressement consiste à choisir judicieusement les variables pour corriger la non-réponse parmi les informations auxiliaires à disposition.

2. Principe du redressement par calage sur marges

La méthode de redressement par repondération la plus utilisée est celle dite de calage sur marges. Selon Roux et Armoogum (2008, 2010), elle « consiste à faire coïncider les marges de quelques variables de l'échantillon à celles de la population cible en modifiant la pondération. Lorsque les variables auxiliaires sont qualitatives, cette approche ne nécessite pas la connaissance dans la population du croisement de ces variables auxiliaires. »

L'idée générale de la méthode d'estimation par calage a été développée et formalisée par Deville et Särndal (1992), bien que de nombreux anciens travaux utilisaient déjà les méthodes d'ajustements de tableaux à des marges connues (Deming et Stephan, 1940 ; Lemel, 1976 ; Madre, 1979, 1980 ; Stephan, 1942).

2.1. Technique de calage

Soit une population finie $U = \{1, ..., k, ... N\}$ dans laquelle un échantillon s est tiré selon un plan de sondage donné $p(.)$. La quantité $p(s)$ représente la probabilité qu'un échantillon aléatoire S prend comme valeur l'échantillon s, i.e. $Pr(S=s) = p(s)$. La probabilité d'inclusion d'ordre 1 de l'unité k dans l'échantillon est notée π_k. De même, la probabilité d'inclusion d'ordre 2 des unités k et l dans l'échantillon est notée π_{kl}. Ces probabilités sont supposées strictement positives.

Soit $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ})^T$ un vecteur de caractères auxiliaires de l'unité *k*. Le vecteur des totaux de **x** dans la population, noté $\mathbf{t}_x$, est supposé connu. Soit y_k la valeur de la variable d'intérêt pour l'unité *k*. L'objectif est d'estimer le total de la variable d'intérêt **y**

$$\mathbf{t}_y = \sum_{k \in U} y_k.$$

En considérant l'information auxiliaire caractérisée par le vecteur $\mathbf{t}_x$ de totaux connus, l'estimateur par calage du total de la variable **y** s'écrit comme :

$$\hat{\mathbf{t}}_{y,w} = \sum_{k \in s} w_k y_k.$$

Évidemment, les poids w_k dépendent de l'échantillon *s* et satisfont l'équation de calage :

$$\sum_{k \in s} w_k \mathbf{x}_k^T$$

où les poids w_k doivent être proches des poids $d_k = 1/\pi_k$.
La proximité entre les poids w_k et $1/\pi_k$ est définie en utilisant une pseudo-distance notée $G_k(.\ ,.)$ supposée définie positive, dérivable et strictement convexe par rapport à w_k. Les poids w_k sont obtenus en minimisant la quantité

$$\sum_{k \in s} G_k(w_k, d_k)$$

sous la contrainte de l'équation de calage.

Plusieurs distances peuvent être utilisées et sont discutées par Deville et Särndal (1992). En général, les poids w_k s'obtiennent en résolvant en $\boldsymbol{\lambda}$, au moyen de la méthode de Newton, le système d'équation :

$$\mathbf{t}_x = \sum_{k \in s} d_k \mathbf{x}_k^T F(\mathbf{x}_k \boldsymbol{\lambda}),$$

$\boldsymbol{\lambda}$ représentant le vecteur des J multiplicateurs de Lagrange. Finalement,

$$w_k = d_k F(\mathbf{x}_k \boldsymbol{\lambda}),$$

F(.) représentant l'inverse de la fonction $g_k(w_k,d_k)$ qui est la dérivée de $G_k(w_k,d_k)$ par rapport à w_k.

L'estimateur par la régression est un cas particulier de l'estimateur par calage, où la pseudo-distance est de type linéaire et est définie comme suit

$$G_k(w_k, d_k) = \frac{(w_k - d_k)^2}{2d_k}.$$

Les poids de calage obtenus avec cette pseudo-distance peuvent prendre des valeurs négatives.

Dans ce papier, nous utiliserons la pseudo-distance de type logistique afin de ne pas obtenir des poids ni trop élevés, ni négatifs. En effet, en considérant deux bornes strictement positives *L* et *H*, la pseudo-distance est donnée par

$$G_k(w_k, d_k) = (a_k \log \frac{a_k}{1-L} + b_k \log b_k H - 1)\frac{1}{A} \quad \text{si} \quad L d_k < w_k < H d_k$$
$$= \infty \qquad \text{sinon,}$$

où $a_k = \frac{w_k}{d_k} - L$, $b_k = H - \frac{w_k}{d_k}$ et $A = \frac{H-L}{(1-L)(H-1)}$. Les bornes *L* et *H* sont choisies arbitrairement de manière à pouvoir réaliser un calage avec un intervalle *[L ; H]* le plus petit possible. Le choix de l'intervalle *[L ; H]* permet de limiter les valeurs poids w_k, et donc d'éviter des poids trop élevés. Malgré cela, cette méthode peut cependant fournir des poids élevés lorsque les variables auxiliaires sont catégorielles, le nombre de contraintes de calage augmente.

Un des avantages du calage est que dès lors que nous disposons de variables auxiliaires corrélées avec la variable d'intérêt et les variables expliquant le mécanisme de réponse, l'estimateur est asymptotiquement sans biais (voir, Deville et Särndal, 1992).

2.2. Estimation de la variance d'un estimateur calé

L'estimateur par calage peut être vu comme un estimateur par la régression. L'estimateur par la régression en est d'ailleurs un cas particulier (voir Deville et Särndal, 1992). Deville et Särndal (1992) ajoute même que tous les estimateurs par calage, quelles que soient leurs pseudo-distances, sont asymptotiquement équivalents.

Tout comme pour l'estimateur par la régression, l'estimation de la variance d'un estimateur par calage peut donc s'obtenir par la technique de linéarisation (voir, par exemple, Tillé, 2001). Deville et Särndal (1992) et Deville et al (1993) ont montré que :

$$\begin{aligned} \mathrm{AVar}(\hat{t}_{y,w}) &\simeq \mathrm{Var}(\hat{t}_E) \\ &\simeq \sum_{k\in U}\sum_{l\in U}\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}\Delta_{kl} \end{aligned}$$

où $E_k = y_k - \mathbf{x}_k^T\mathbf{B}$ donne les résidus de la régression de **y** sur le jeu des variables auxiliaires **x** au niveau de la population.

Une approximation de la variance est alors donnée par :

$$\hat{\mathrm{Var}}(\hat{t}_{y,w}) = \sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}w_k e_k w_l e_l,$$

où $e_k = y_k - \mathbf{x}_k^T\hat{\mathbf{B}}_{ws}$ donne les résidus de la régression *w*-pondérée de **y** sur le jeu des variables auxiliaires **x** au niveau de l'échantillon.

Les poids de calage w_k étant calculés de manière à être très proches des poids de sondage d_k, Deville et Särndal (1992) ont montré que l'estimateur par calage est asymptotiquement sans biais, l'estimateur d'Horvitz et Thompson (1952) étant sans biais. De plus, la variance de l'estimateur par calage est d'autant plus faible dès lors que les variables auxiliaires sont très corrélées avec la variable d'intérêt.

Puisque l'estimateur de la variance se calcule à partir des résidus de la régression *w*-pondérée de **y** sur le jeu des variables auxiliaires **x**, il est plus petit que la variance de l'estimateur d'Horvitz et Thompson (1952).

2.3. Comment réduire la variance ?

D'après El Haj Tirari (2012), lorsque nous utilisons un grand nombre de variables auxiliaires, la variance de notre estimateur peut augmenter. Car pour minimiser la variance, il faut minimiser la somme du produit des poids de calage et des résidus. L'introduction de l'information auxiliaire dans le calage permet de diminuer les résidus mais augmente aussi la dispersion des poids. Il faut donc choisir judicieusement les variables auxiliaires qui permettent de diminuer la variance lors du calage pour le redressement d'une enquête. En effet, il n'est pas nécessaire de mettre toutes les variables auxiliaires dans un calage surtout lorsque ces variables sont corrélées entre elles. Le principal but de notre article, consiste à choisir l'information auxiliaire pour une enquête de mobilité. Nous verrons par la suite que le fait de réduire le nombre de variables auxiliaires permettra d'améliorer l'estimation du nombre de voitures dans une région de France.

3. L'Enquête Nationale sur les Transports et les Déplacements (ENTD) 2007-2008

3.1. Présentation de l'ENTD

L'Institut National de la Statistique et des Études Economiques (INSEE) présente l'ENTD 2007-2008 comme suit :

« Tous les dix ans environ, le ministère chargé des Transports, l'INSEE et l'Institut National de Recherche sur les Transports et leur Sécurité[1] (INRETS) conduisent une Enquête Nationale sur les Transports (ENTD). L'ENTD 2007-2008 succède à celle de 1993-1994 et les précédentes enquêtes datent de 1966-67, 1973-74 et 1981-82. L'objectif de ces enquêtes est la connaissance des déplacements des ménages résidant en France et de leur usage des moyens de transport tant collectifs qu'individuels. Elle permet d'avoir une vision globale et cohérente de la mobilité et d'analyser le parc de véhicules dont disposent les ménages et de leur usage. Elle permet aussi de répondre aux questions sur les trafics interrégionaux et internationaux dont les enjeux sont très importants en matière d'investissements et de mesurer les distances parcourues dont la connaissance est indispensable pour appréhender les problématiques environnementales. Par rapprochement avec les résultats des enquêtes précédentes, elle permet des comparaisons dans le temps et dans l'espace. »

3.2. Redressement de l'ENTD

L'échantillon de l'ENTD a été tiré à partir de l'Echantillon Maître de 1999 (EM 99) de l'INSEE, qui lui-même a été tiré à partir du recensement de la population de 1999. Le recensement de 1999 est donc une source d'information auxiliaire complète permettant d'analyser le mécanisme de réponse. Les variables susceptibles d'expliquer le mécanisme de réponse sont disponibles pour tout l'échantillon de l'ENTD (répondants et non-répondants). Armoogum et Roux (2012) ont mis en évidence les variables auxiliaires qui permettent d'expliquer le mécanisme de réponse de l'ENTD au moyen d'un modèle logistique (voir Tableau 1).

Tableau 1 – *Liste des variables disponibles dans la base de sondage et analyse du mécanisme de réponse pour l'ENTD*

Variable disponible dans la base de sondage	Variable expliquant le mécanisme de réponse
Zone de résidence	X
Revenu moyen de la commune du logement au RP99	
Tranche d'unité urbaine de la commune en 1999	
Période d'achèvement de l'immeuble	

[1] Depuis le 1er janvier 2011, l'Institut National de Recherche sur les Transports et leur Sécurité (INRETS) et le Laboratoire Central des Ponts et Chaussées (LCPC) ont fusionné pour donner naissance à l'Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR).

Appartenance de l'immeuble à un organisme d'habitation à loyer modéré (HLM)	X
Type de bâtiment	X
Immeuble disposant d'un ascenseur	
Nombre de pièces du logement	X
Surface du logement	
Nombre de personnes du ménage au RP99	
Age de la personne de référence au RP99	X
Sexe de la personne de référence au RP99	
Catégorie socio-professionnelle de la personne de référence au RP99	
Motorisation du ménage au RP99	X
Vague de l'enquête	X

Source : INSEE, SOES, IFSTTAR : ENTD 2007–2008.

Selon Armoogum et Roux (2012), le mécanisme de réponse pour l'ENTD oppose en première analyse :

- **Type de bâtiment.** Les ménages habitant une maison aux ménages résidant dans une habitation collective. Les échecs sont plus fréquents pour les logements collectifs (c'est probablement une question d'accessibilité du logement) ;
- **Nombre de pièces du logement.** Les ménages habitant un studio ou une chambre aux ménages résidant dans des logements ayant plusieurs pièces. Cette variable est corrélée avec le nombre de personnes vivant dans le ménage. Ainsi, une taille de ménage plus importante s'accompagne d'une probabilité plus grande de réaliser l'entretien ;
- **Zone de résidence.** Les logements situés en zone rurale et en agglomération de moins de 20000 habitants à ceux situés dans l'agglomération de Paris. Les échecs sont d'autant plus nombreux qu'on progresse vers une plus grande urbanisation ;
- **Motorisation du ménage.** Les ménages n'ayant aucune automobile aux ménages motorisés. Les ménages non-équipés en automobile sont moins favorables à la réalisation des entretiens ;
- **Age de la personne de référence.** Les ménages dont la personne de référence a moins de 35 ans ou plus de 65 ans à ceux dont l'âge se situe entre 35 et 65 ans. Certainement pour des raisons différentes, les taux d'échec sont plus importants pour les ménages dont la personne de référence a moins de 35 ans et pour ceux dont l'âge de la personne de référence est supérieur à 65 ans. Pour les premiers cela souligne la difficulté des enquêteurs de joindre ces ménages et pour les seconds la réticence des personnes âgées à répondre à un long questionnaire ;
- **Appartenance de l'immeuble à un organisme d'habitation à loyer modéré (HLM).** Les ménages résidants dans une HLM aux autres. Les échecs sont plus nombreux pour les ménages habitant une HLM ;

- **Vague de l'enquête.** Les ménages interrogés au mois de juillet–août de ceux interrogés à un autre moment de l'année. Les échecs sont plus nombreux pendant les vacances d'été, période au cours de laquelle nous supposons que les ménages sont les plus mobiles.

Armoogum et Roux (2012) ont réalisé le redressement de l'ENTD à partir des données disponibles du recensement de 2008, en utilisant au maximum les variables qui expliquent le mécanisme de réponse et les variables auxiliaires corrélées avec la mobilité (voir Tableau 2).

Tableau 2 – *Liste des variables utilisées pour le redressement de l'ENTD*

	Variable expliquant le mécanisme de réponse	**Variable utilisée pour le redressement de l'ENTD**
Zone de résidence	X	X
Appartenance de l'immeuble à un organisme d'habitation à loyer modéré (HLM)	X	
Type de bâtiment	X	X
Nombre de pièces du logement	X	
Age de la personne de référence	X	
Motorisation du ménage	X	X
Vague de l'enquête	X	X
Catégorie socio-professionnelle de la personne de référence		X
Personne de référence par sexe x âge		X
Type du ménage		X
Nationalité de la personne de référence		X
Nombre d'individus par sexe x âge		

3.3. Estimations rhônalpines à partir du redressement national

Des premières estimations régionales peuvent être obtenues à partir du redressement national de l'ENTD. Le Tableau 3 donne les estimations du nombre total de voitures, du nombre de voitures fonctionnant au diesel et du nombre de voitures fonctionnant à l'essence et autres, par ménage, au niveau de la région Rhône-Alpes ainsi que les erreurs relatives et écarts-types associés. Les erreurs relatives fournies sont obtenues par le produit des coefficients de variation avec le quantile d'ordre 2,5% de la loi normale (soit 1,96). Ces estimations sont issues du redressement au niveau national

de l'ENTD 2007-2008. Pour le calcul des variances, le plan de sondage de l'échantillon national est approché par un plan de Poisson. Comme le souligne Le Guennec (2012), ceci est dû au problème d'« accès à tous les paramètres du tirage de l'échantillon national » (paramètres de tirage de l'EM 99 à partir du recensement de la population de 1999, paramètres de tirage de l'ENTD à partir de l'EM 99). De même, le nombre de ménages en Rhône-Alpes est supposé connu (à partir du recensement de la population de 2008).

Tableau 3 – *Estimation du parc de voiture en Rhône-Alpes avec redressement au niveau national de l'ENTD 2007–2008*

	Estimation	
	Moyenne (± Erreur Relative)	**Ecart-type**
Nombre total de voitures	1,32 (± 6,76%)	0,045
Nombre de voitures diesel	0,70 (± 8,72%)	0,031
Nombre de voitures essence et autres	0,62 (± 8,94%)	0,028

Source : INSEE, SOES, IFSTTAR : ENTD 2007–2008.

Le redressement de l'ENTD ayant été réalisé au niveau national, celui-ci peut ne pas tenir compte des spécificités des régions. Il est donc préférable d'effectuer un nouveau redressement au niveau de la région Rhône-Alpes. Les données régionales du recensement de la population de 2008 étant disponibles, les estimations au niveau de la région Rhône-Alpes peuvent être améliorées en calant directement sur la région.

Dans la pratique, lorsque la taille de l'échantillon est suffisamment grande, il est assez « facile » de satisfaire aux équations de calage. Mais, plus la taille de l'échantillon est faible, plus la précision des estimations par calage risque de diminuer à cause des fortes contraintes de calage. Dans la suite de ce papier, nous nous restreindrons au redressement du sous-échantillon rhônalpin.

3.4. Le sous-échantillon rhônalpin

L'échantillon des ménages répondants de l'ENTD 2007-2008 compte 20178 ménages sur toute la France et 986 ménages au niveau de la région Rhône-Alpes. On souhaite estimer, par ménage (l'unité statistique), le nombre total de voitures particulières, le nombre de voitures fonctionnant au diesel, et de voitures fonctionnant à l'essence et autres, au niveau de cette région ainsi que les précisions associées.

Les trois variables d'intérêt sont liées entre elles car le nombre de voitures particulières est égale à la somme du nombre de voiture diesel et du nombre de voiture essence. Toutefois, lorsque nous regardons les coefficients de corrélation, nous observons (voir Tableau 4) :

- Les trois variables d'intérêt considérées ne sont pas forcément très corrélées entre elles. Les coefficients de corrélation les plus élevés sont toujours pour le nombre total de voitures avec les nombres de voitures fonctionnant au diesel ou essence. Le coefficient de corrélation entre le nombre de voiture fonctionnant au diesel et le nombre de voiture fonctionnant à l'essence reste faible.

- Les coefficients de corrélation au niveau de la région Rhône-Alpes sont beaucoup plus faibles que ceux de la France entière. La corrélation entre le nombre de voiture fonctionnant au diesel et le nombre de voiture fonctionnant à l'essence est négative.

Ceci suggère une certaine spécificité de la région Rhône-Alpes avec le reste de la France. Les ménages multi-motorisés en France ont plus souvent un parc de voitures composé de véhicules fonctionnant à l'essence et au diesel, que les ménages de la région Rhône-Alpes. Cela laisse supposer que les comportements des ménages, en terme d'équipement en voiture, sont très différents. Il est donc légitime d'effectuer un redressement direct du sous-échantillon à partir des données du recensement de 2008 pour la région Rhône-Alpes.

Tableau 4 – *Matrices des corrélations entre les trois variables d'intérêt considérées, au niveau national et au niveau de la région Rhône-Alpes*

		Nombre total de voitures	**Nombre de voitures diesel**	**Nombre de voitures essence et autres**
Nombre total de voitures	*National*	1	0,82	0,76
	Rhône-Alpes	1	0,67	0,53
Nombre de voitures diesel	*National*	0,82	1	0,24
	Rhône-Alpes	0,67	1	-0,27
Nombre de voitures essence et autres	*National*	0,76	0,24	1
	Rhône-Alpes	0,53	-0,27	1

Source : INSEE, SOES, IFSTTAR : ENTD 2007–2008.

3. Redressement rhônalpin

A partir du recensement de la population de 2008, nous disposons de plusieurs marges connues au niveau de la région Rhône-Alpes. Nous souhaitons savoir quelles sont les variables auxiliaires qui peuvent améliorer nos estimateurs. Ces marges disponibles sont les suivantes : la **motorisation** (ménage sans voiture ; ménage ayant au moins une voiture), le **type du ménage**, l'**âge** de la personne de référence du ménage, le **sexe** de la personne de référence du ménage, la **zone de résidence**, le **type d'aire urbaine de résidence**, le **type du logement**, la **taille du ménage**, la **catégorie socio-professionnelle** de la personne de référence du ménage et la **vague de l'enquête**.

La variable auxiliaire **motorisation** est intuitivement très corrélée à la variable d'intérêt **nombre de voitures** mais ces deux variables sont totalement distinctes et ne sont pas les mêmes. La variable auxiliaire **motorisation** fournit l'information : nombre de ménages n'ayant aucun véhicule ; nombre de ménages ayant au moins un véhicule. C'est cette information qui est fournie par le recensement de 2008.

Notons qu'au départ, nous nous basons sur les variables qui corrigent la non-réponse et les variables qui sont corrélées avec les variables d'intérêt pour réaliser les estimations. Réaliser un calage avec toutes ces informations auxiliaires ne donnerait pas automatiquement une estimation avec la meilleure précision, à cause notamment des fortes contraintes de calage à satisfaire. Un choix judicieux des variables uniquement utiles doit être fait pour améliorer la précision de nos estimateurs. Nous proposons une procédure qui permet d'identifier les variables auxiliaires à utiliser afin d'obtenir des estimations par calage avec des précisions optimales. La procédure est intimement similaire à une régression pas à pas : les différences résident dans l'utilisation du calage et de calculs de variance pour sélectionner les variables pertinentes.

3.1. Principe du choix des variables pertinentes pour atteindre les précisions optimales

La procédure de sélection des variables se fait en deux étapes. La première consiste à éliminer les variables auxiliaires non significatives, en ayant recours au critère de l'AIC[2]. En principe, après cette première étape, la variance de l'estimateur calé sur les variables retenues comme significatives devrait être minimale. Cependant, les poids de calage obtenus, utilisés dans le calcul de variance, sont très instables. Les poids initiaux de calage sont très dispersés et peuvent prendre des valeurs très élevées. De plus, le calage est effectué avec des variables catégorielles. Les poids finaux de calage sont en conséquence très dispersés à leur tour et peuvent également prendre des valeurs très élevées malgré l'utilisation de la pseudo-distance de type logistique. En enlevant d'autres variables auxiliaires dans la procédure de calage, les variances peuvent donc encore diminuer. La deuxième étape de la procédure de sélection des variables intervient dans ce cadre. En notant p le nombre de variables auxiliaires retenues comme significatives par le critère de l'AIC, la deuxième étape de la procédure consiste à calculer p variances en n'utilisant dans les calages que p - 1 variables sur les p à chaque fois, chacune des p variables étant mise de côté une seule fois. Ainsi, les p - 1 variables associées à la plus petite variance sont retenues si cette nouvelle variance est inférieure à celle obtenue avec les p variables. La procédure est ensuite répétée en calculant p - 1 variances en n'utilisant dans les calages que p - 2 variables sur les p - 1 à chaque fois, chacune des p - 1 variables étant mise de côté une seule fois. Les p - 2 variables associées à la plus petite variance sont retenues si cette nouvelle variance est inférieure à celle obtenue avec les p - 1 variables. Et ainsi de suite. Dans le cas où la nouvelle variance n'est pas inférieure à celle obtenue précédemment, deux variables sont simultanément mises de côté, puis si nécessaire trois variables simultanément, . . ., jusqu'à p - 1 variables simultanément si nécessaire.

3.2. Algorithme de sélection des variables auxiliaires pertinentes

L'algorithme suivant est proposé afin de sélectionner les variables auxiliaires pertinentes pour le redressement du sous-échantillon de la région Rhône-Alpes.

1. Considérer les 10 variables de calage et calculer la variance de l'estimateur obtenu par calage sur les 10 variables.
2. Par le critère d'Akaike (AIC), déterminer les variables considérées comme non significatives et voir l'ordre de non-significativité des variables.

[2] Le critère d'Akaike (AIC) est défini par la formule : AIC = 2k - 2 ln L où k est le nombre de paramètres dans le modèle considéré et L est la fonction de vraisemblance.

3. Tant que la variance diminue

    i. Enlever la variable la moins significative des variables considérées comme non significatives et calculer la variance de l'estimateur obtenu par calage sur les variables restantes

4. Considérer les variables de calage restantes.

5. Tant que la variance diminue

    i. Retirer à chaque fois une variable et calculer la variance associée à l'estimateur obtenu.

    ii. Considérer les variables de calage restantes ayant la plus petite variance et étant inférieure à la variance du précédent modèle.

6. Si la variance ne diminue pas, refaire 5 mais en retirant deux variables simultanément, puis si nécessaire, trois variables simultanément, puis quatre variables, …

3.3. Résultats de la procédure de sélection des variables

Pour chacune des variables d'intérêt considérées, **nombre total de voitures**, **nombre de voitures diesel** et **nombre de voitures essence et autres**, le Tableau 5 résume la comparaison entre les précisions minimales obtenues par la procédure de sélection et les précisions obtenues par calage global sur toutes les variables auxiliaires disponibles au niveau de la région Rhône-Alpes. Le calage global est le redressement qui corrige au maximum les différents biais dus à l'échantillonnage et à la phase de non-réponse. Nous notons que les écarts relatifs entre les deux estimations sont faibles (0,7% pour le nombre total de voiture ; 1,3% pour le nombre de voitures diesel ; et 3,3% pour le nombre de voitures essence) et que les intervalles de confiance se chevauchent très largement.

Tableau 5 – *Comparaison entre les précisions optimales obtenues par la procédure de sélection et les précisions obtenues par calage global sur toutes les variables auxiliaires disponibles au niveau de la région Rhône-Alpes*

	Redressement régional			
	Redressement Global		**Redressement Optimal**	
	Moyenne (± Erreur Relative)	**Ecart-type**	**Moyenne (± Erreur Relative)**	**Ecart-type**
Nombre total de voitures	1,36 (± 3,32%)	0,023	1,35 (± 2,98%)	0,020
Nombre de voitures diesel	0,75 (± 6,66%)	0,025	0,74 (± 6,27%)	0,024
Nombre de voitures essence et autres	0,61 (± 7,54%)	0,023	0,63 (± 6,67%)	0,021

Source : INSEE, SOES, IFSTTAR : ENTD 2007–2008.

Tableau 6 – *Variables auxiliaires sélectionnées pour l'optimisation des variances*

	Nombre total de voitures	Nombre de voitures diesel	Nombre de voitures essence et autres
Motorisation	X	X	X
Type du ménage	X	X	
Sexe		X	
Age	X	X	
Zone de résidence			
Type de l'aire urbaine			
Type du logement	X	X	X
Taille du ménage	X	X	
Catégorie socio-professionnelle		X	
Vague		X	

Source : INSEE, SOES, IFSTTAR : ENTD 2007–2008.

Le Tableau 6 présente un récapitulatif des différentes variables de calage utilisées pour obtenir les précisions minimales pour chacune des variables d'intérêt considérées.

En comparant le Tableau 3 et le Tableau 5, nous pouvons clairement constater une nette amélioration de la précision entre les estimations nationales et les estimations par calage direct au niveau de la région Rhône-Alpes. La précision a augmenté de 49% pour la variable **nombre total de voitures** en faisant un simple calage sur toutes les variables auxiliaires disponibles au niveau de la région. Pour les variables **nombre de voitures diesel** et **de voitures essences et autres**, la précision a augmenté respectivement de 18 et 17%. Comme ces deux dernières variables se focalisent sur des domaines d'estimation encore plus restreints, l'augmentation de la précision est moindre comparée à la précision de l'estimation de la variable **nombre total de voitures**.

Le Tableau 5 montre également que le redressement au niveau de la région avec toutes les variables auxiliaires peut encore être amélioré. Les précisions optimales des estimations sont obtenues avec un nombre plus petit de variables auxiliaires. Ceci s'explique par le fait qu'en redressant au niveau de la région (la taille de l'échantillon considéré est alors réduite car nous travaillons au niveau d'une sous-population, la région Rhône-Alpes, et non plus au niveau de la population entière, la France entière), redresser avec moins de variables permet de relâcher les contraintes de calage. Relâcher les contraintes augmente le nombre de degrés de liberté et permet d'avoir des poids de calage moins dispersés. Ainsi, par ce relâchement de contrainte, la précision optimale est de 11% meilleure qu'avec un redressement avec toutes les variables auxiliaires, pour la variable d'intérêt **nombre totale de voiture**. Les précisions optimales pour les variables **nombre de voitures diesel** et **de voitures essence et autres** augmentent respectivement de 7 et 8% en comparaison d'un redressement au niveau de la région avec toutes les variables auxiliaires.

3.4. Redressement régional avec un système unique de pondération

Par la procédure de sélection de variables auxiliaires pertinentes, le Tableau 6 nous montre que les précisions optimales sont obtenues avec des jeux différents de variables auxiliaires pour chacune des variables d'intérêt considérées. Cependant, en sondage, il est plutôt d'usage de faire appel à un système unique de pondérations afin de fournir des estimations pour différentes variables d'intérêt. En effet, dans les grandes enquêtes regroupant quelques dizaines, voire des centaines, de variables d'intérêt, il serait très fastidieux, mais non impossible, de lancer un redressement optimal pour chacune des variables d'intérêt. Il est donc judicieux de ne considérer qu'un système unique de pondérations. Le choix des variables auxiliaires permettant d'obtenir le système unique de poids doit être fait, en fonction du sujet d'estimation, de telle sorte à ne diminuer que sensiblement la précision.

Dans ce papier, nous nous intéressons à la motorisation des ménages, notamment le nombre total de voitures, le nombre de voitures fonctionnant au diesel, et le nombre de voitures fonctionnant à l'essence et autres. La pondération issue du redressement optimal de la variable **nombre total de voitures** est un bon compromis.

Le Tableau 7 donne une comparaison des précisions entre les redressements optimaux et le redressement par système unique de pondération. La perte de précision pour la variable **nombre de voitures diesel** n'est que de 0,9% et celle de la variable **nombre de voitures essence et autres** est de 3,9%.

Tableau 7 – *Comparaison entre les précisions minimales obtenues par la procédure de sélection et les précisions obtenues avec un système unique de pondération*

	Redressement régional			
	Redressement Optimal		***Système Unique de Pondération***	
	Moyenne (± Erreur Relative)	**Ecart-type**	**Moyenne (± Erreur Relative)**	**Ecart-type**
Nombre total de voitures	1,35 (± 2,98%)	0,020	1,35 (± 2,98%)	0,020
Nombre de voitures diesel	0,74 (± 6,27%)	0,024	0,72 (± 6,50%)	0,024
Nombre de voitures essence et autres	0,63 (± 6,67%)	0,021	0,63 (± 6,89%)	0,022

Source : INSEE, SOES, IFSTTAR : ENTD 2007–2008.

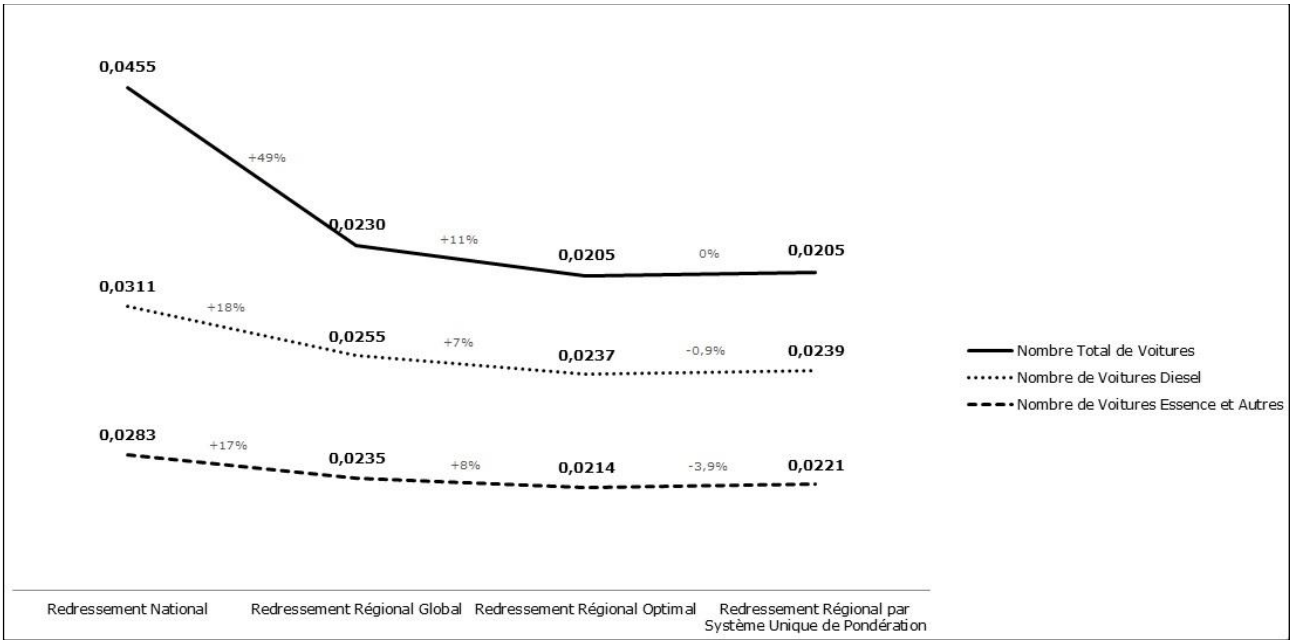

Figure 1 – *Evolution des écarts-types des estimations suivant les différents redressements utilisés*

Conclusion

La méthode présentée dans ce papier dépend des variables auxiliaires disponibles ainsi que de leur pouvoir explicatif sur les variables d'intérêt considérées. Elle permet de sélectionner les bonnes variables auxiliaires à choisir pour le redressement d'un sous-échantillon. Dans le cas du sous-échantillon de la région Rhône-Alpes de l'ENTD 2007–2008, les précisions minimales obtenues pour le nombre total de voitures particulières, le nombre de voitures utilisant du diesel, et de voitures utilisant de l'essence et autres, dépendent clairement du choix des variables auxiliaires utilisées dans les procédures de calage. Afin d'éviter différentes pondérations distinctes pour chacune de ces variables d'intérêt, un système unique de pondérations a été établi. Les précisions finales obtenues avec ce système unique de pondérations ont la caractéristique d'être assez équivalentes avec les précisions minimales résultant de la procédure de sélection des variables auxiliaires pertinentes pour chacune des variables d'intérêt.

La méthode proposée dans ce papier peut être transférée à d'autres variables d'intérêt d'un autre thème : il « suffit» pour cela de retrouver les bonnes variables auxiliaires à utiliser pour le redressement. Ces variables auxiliaires peuvent par exemple être une combinaison de variables socio-démographiques et de variables qui expliquent le mécanisme de réponse.

Dans le cas où l'on serait amené à travailler avec un nombre trop important de variables d'intérêt d'un même thème, il est possible de choisir parmi ces variables, un nombre restreint de variables d'intérêt. On pourra ainsi appliquer la méthode proposée dans cet article à ce nombre restreint de variables d'intérêt. Le système unique de pondérations ainsi obtenu pourra être ensuite utilisé pour l'ensemble de toutes les variables d'intérêt.

La méthode utilisée rentre dans le cadre de l'estimation sur petits domaines (méthodologie d'estimation largement développée par Rao, 2003, 2015) où dans notre cas, le domaine est le sous-échantillon rhônalpin. La méthode proposée est dite directe car elle ne fait intervenir que

l'information disponible au niveau du domaine. L'inconvénient de la méthode présentée est que la précision est faible dès lors que la taille du sous-échantillon, inclus dans le domaine, est trop petite. La méthode présentée est donc fortement dépendante de la taille du sous-échantillon d'étude.
Afin de pallier à cette limite de la méthode proposée, il est sans doute nécessaire d'emprunter de la « force » en dehors du domaine considéré. On parle alors de méthode indirecte d'estimation sur petits domaines.

Références

Armoogum J (2000) Correction de la non-réponse et de certaines erreurs de mesures dans une enquête par sondage : application à l'Enquête sur les Transports et Communications de 1993-1994, Thèse de doctorat, Université Libre de Bruxelles

Armoogum J, Roux S (2012) Mise en perspective des Enquêtes Nationales Transports 1973/74 -- 1981/82-- 1993/94 -- 2007/08, Note méthodologique, IFSTTAR

Beaumont, JF (2005) On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment." Survey Methodology 31(2): 227-231.

Chauvet G, Goga C (2012). Redresser un échantillon ... mais pas trop. Notes de cours. 44es Journées de Statistique, Bruxelles

Deming WE, Stephan FF (1940) On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. The Annals of Mathematical Statistics, Institute of Mathematical Statistics 11:427-44

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. Journal of the American Statistical Association 87:376-82

Deville JC, Särndal CE, Sautory O (1993) Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association 88:1013-20

El Haj Tirari M (2012), Critère du choix des variables auxiliaires à utiliser dans l'estimateur par calage, 7e colloque francophone sur les sondages, Rennes

Emrich L (1983), Randomized response techniques, dans Incomplete Data in Sample Surveys, Madow WG, Olkin I, et Rubin, DB Eds. Vol II: Theory and Bibliographies New York: Academic Press, pp. 73-80

Goga, C, Shehzad, M-A and Vanheuverzwyn, A (2011) Principal Component Regression with Survey Data. Application on the French Media Audience, Proceedings of the 58th ISI World Statistics Congress, Dublin

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47*:* 663-85

Le Guennec J (2012) Application de méthodes « petits domaines » à des estimations régionales dans l'Enquête Nationale sur les Transports et les Déplacements 2007-2008. Actes des Journées de Méthodologie Statistique 2012

Le Guennec J, Sautory O (2002) Application du calage généralisé à la correction de la non-réponse : une expérimentation. Actes des Journées de Méthodologie Statistique 2002

Lemel Y (1976) Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondages. Annales de l'Insee 273-81

Madre JL (1979) Ajustement et extrapolation de tableaux statistiques. Thèse de doctorat, Université Pierre et Marie Curie

Madre JL (1980) Méthode d'ajustement d'un tableau à des marges. Les cahiers de l'Analyse des données 87-99

Rao JNK (2003) Small Area Estimation. New York: Wiley

Rao JNK, Molina I (2015) Small Area Estimation. New York: Wiley

Rao JNK, Singh AC (2009) Range Restricted Weight Calibration for Survey Data Using Ridge Regression. Pakistan Journal of Statistics 25(4): 371-384.

Razafindranovona T (2015) La collecte multimode et le paradigme de l'erreur d'enquête totale, Série des documents de travail « Méthodologie Statistique », de la Direction de la Méthodologie et de la Coordination Statistique et Internationale de l'INSEE, M 2015/01

Roux S (2012) Transition de la motorisation en France au 20e siècle. Thèse de doctorat, Université Paris-Sorbonne

Roux S, Armoogum J (2008) Correction de la non-réponse dans l'Enquête Nationale sur les Transports et les Déplacements 2007-2008, Note méthodologique, Département Économie et Sociologie des Transports - INRETS

Roux S, Armoogum J (2010) Redressement de l'Enquête Nationale sur les Transports et les Déplacements 2007-2008 Note méthodologique, Département Économie et Sociologie des Transports - INRETS

Roux S, Armoogum J (2011) Calibration Strategies to Correct Nonresponse in a National Travel Survey. Transportation Research Records: Journal of the Transportation Research Board *2246*:1-7

Särndal C, Swensson B (1987) A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse. International Statistical Review / Revue Internationale De Statistique, 55(3): 279-294.

Stephan FF (1942) An Iterative Method of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known. The Annals of Mathematical Statistics, Institute of Mathematical Statistics 13:166-78

Tillé Y (1992) Utilisation a posteriori d'informations auxiliaires en théorie des sondages sans référence à un modèle. Thèse de doctorat, Université Libre de Bruxelles

Tillé Y (2001) Théorie des sondages : Échantillonnage et estimation en populations finies. Éditions Dunod, 284 p