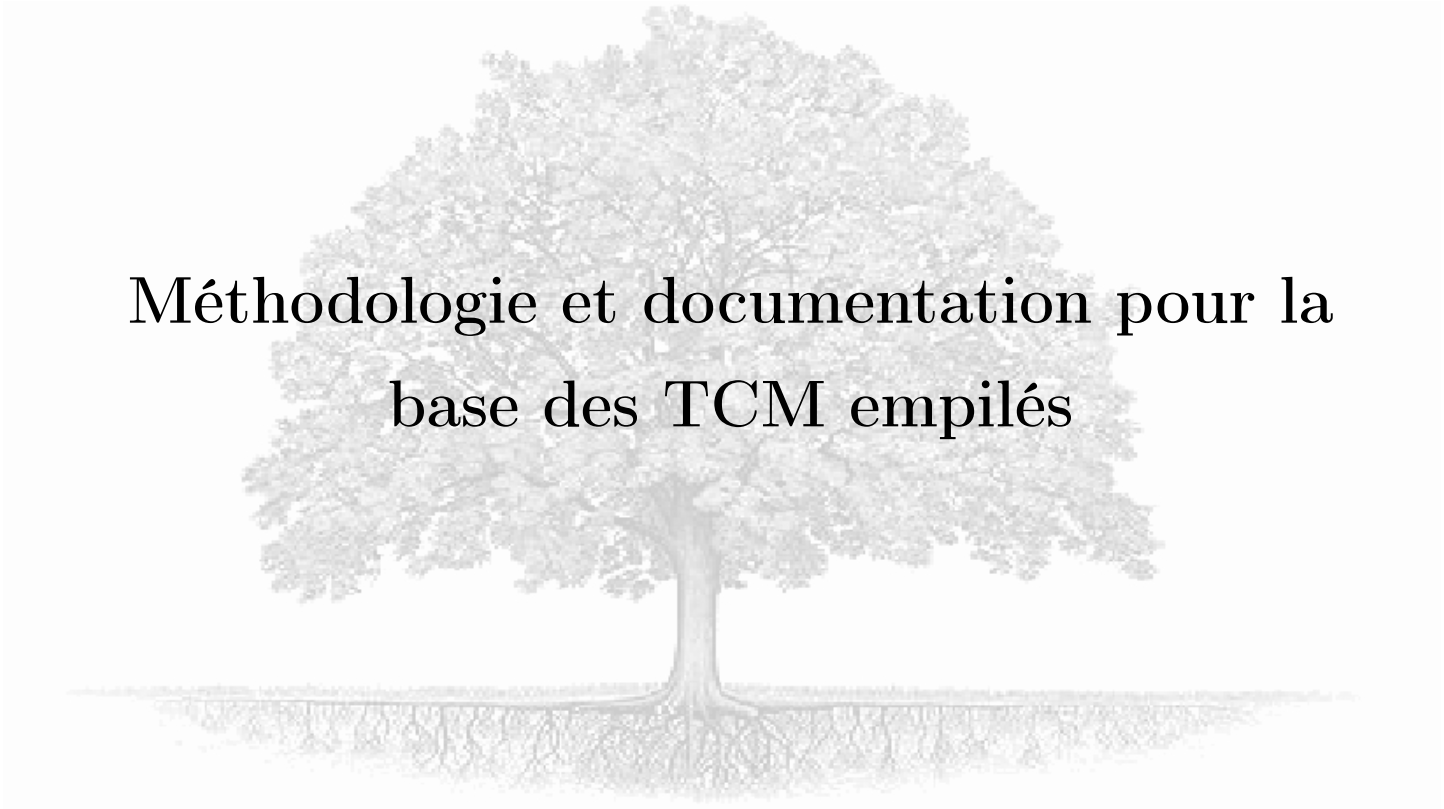
INSEE



Méthodologie et documentation pour la base des TCM empilés

Céline Leroy

Août 2018

Table des matières

Introduction

Le Tronc Commun des enquêtes Ménages (TCM) a été mis en place en 2004 en vue d'harmoniser le recueil des informations communes aux enquêtes ménages de l'Insee, mais également de manière à standardiser les procédures de traitement informatique, notamment la conception des tables TCM [1] à l'aval de l'enquête.

Le TCM permet de recueillir des informations très riches sur les habitants du logement, leurs relations familiales, la mise en commun des dépenses et des ressources, la multi-résidence [2], etc. Ces mêmes informations étant recueillies pour toutes les enquêtes ménages embarquant le TCM, ce dernier pourrait devenir un puits d'information précieux si l'on rassemble les données issues des différentes enquêtes.

Par ailleurs, certains chercheurs de l'Institut national des études démographiques (Ined), ainsi que la division Enquêtes et Études démographiques de l'Insee sont intéressés par l'exploitation des situations familiales rares comme les familles monoparentales, les familles recomposées ou encore les situations de multi-résidence.

L'idée d'empiler les bases de données issues du TCM de chaque enquête prend alors tout son sens. En effet, le TCM recueille l'information nécessaire pour ces exploitations mais ces situations familiales sont trop rares pour pouvoir être étudiées par le biais d'une seule enquête. Ainsi, en empilant les bases de données du TCM issues de plusieurs enquêtes sur plusieurs années, il sera possible de faire des études fiables sur ces situations familiales car on disposera de davantage de données.

Un premier empilement, avec un système de pondération non optimal [3], a été réalisé et exploité dans le cadre d'une présentation aux Journées de la Méthodologie Statistique (JMS) en 2012, par un chercheur de l'Ined, Laurent Toulemon, et le responsable du TCM de l'époque, Thomas Denoyelle. L'article consistait à présenter les situations de multi-résidence ainsi que les différentes manières d'en tenir compte, pour l'exploitation des situations familiales rares [1]. C'est alors suite à cette étude que le Cnis (Conseil national de l'information statistique) a fait une demande pour qu'une base des TCM empilés soit diffusée sur le réseau Quetelet [4]. Le Département des Méthodes Statistiques (DMS) a répondu favorablement à cette demande et est donc en charge de produire une base avec un meilleur système de pondération, prenant en compte l'empilement d'échantillons d'enquêtes différentes notamment.

1. Il s'agit des bases de données restreintes aux variables du TCM. Il existe trois tables correspondant à chaque niveau : logement, ménage et individu.

2. La multi-résidence est le fait d'habiter habituellement plusieurs logements (au moins un mois dans l'année en cumulé).

3. Il s'agissait simplement des poids d'enquête fournis par chaque responsable à l'issue des traitements de la non-réponse et du calage.

4. Réseau permettant de mettre des bases de données à disposition des chercheurs en sciences humaines et sociales. Le service des enquêtes de l'Ined est un partenaire de ce réseau.

Néanmoins, le projet de conception de la base des TCM empilés a pris du retard à cause d'autres contraintes sur le TCM, en particulier, la conception d'un TCM adapté à la réinterrogation pour les enquêtes longitudinales. Ainsi, du fait de la difficulté à réaliser cette base des TCM empilés, en attendant sa diffusion *via* le réseau Quetelet, une convention a été signée entre l'Insee et l'Ined pour son exploitation. Le projet n'ayant pas totalement abouti fin août 2018 [5], une base intermédiaire va être mise à disposition des chercheurs *via* la plateforme Big Stat (projet ANR piloté par l'Ined). Elle empile les tables TCM de 27 enquêtes sur la période 2006-2015 [6]. Elle sera utilisée notamment pour des études sur la multi-résidence ; des comparaisons avec d'autres sources, comme le recensement ou l'enquête Famille et Logement, concernant les strutures familiales complexes, ou encore pour l'estimation du nombre d'orphelins et la description de leurs conditions de vie.

Cette note a pour objectif de documenter la source.
Dans un premier temps, toute la méthodologie sur la stratégie de pondération adoptée est présentée. En effet, l'empilement d'échantillons d'enquêtes différentes nécessite un travail de repondération pour garder la représentativité initiale. Il s'agira donc d'expliquer le principe général du partage des poids et de voir dans quelle mesure on peut l'appliquer pour la conception de la base des TCM empilés, du niveau logement au niveau individu [7].
Dans un deuxième temps, figure toute la documentation concernant la base des TCM empilés, à savoir la liste de toutes les enquêtes empilées avec les différents effectifs, le dictionnaire des principales variables [8], ainsi que les spécificités de chaque enquête et les alertes en cas de problème sur le TCM une année donnée.

5. Il faut encore simplifier la lecture des tables, notamment en supprimant les variables de collecte inutiles et en harmonisant certaines variables correspondant à la même notion mais dont les modalités et noms ont évolué entre 2006 et 2015.

6. Parmi ces 27 enquêtes, certaines se répètent, notamment les enquêtes annuelles SRCV (Statistiques sur les Ressources et Conditions de Vie) et CVS (Cadre de Vie et Sécurité).

7. À l'aval de chaque enquête, trois tables restreintes aux variables du TCM sont produites : une table de niveau logement, une table de niveau ménage car un logement peut contenir plusieurs ménages et une table de niveau individu.

8. Ce dernier est à l'état d'ébauche et devra être complété des autres variables secondaires par le successeur de Céline Leroy.

Première partie

Méthodologie pour la conception de la base des TCM empilés

Introduction

La base des TCM empilés contient les bases TCM de 27 enquêtes sur la période 2006-2015.
Cet empilement de différents échantillons, chacun représentatif de la population, nécessite une répondération de manière à garder la représentativité initiale.

Cette partie a donc vocation à expliquer la stratégie adoptée pour la pondération des TCM empilés. On distinguera trois niveaux différents et deux types d'empilement pour les tables TCM. Les troix niveaux correspondent aux niveaux logement, ménage [9] et individu. Quant aux types d'empilement, il y a, d'une part, l'empilement annuel qui consiste à empiler les tables TCM des enquêtes d'une même année et, d'autre part, l'empilement périodique qui consiste à empiler les tables annuelles sur la période 2006-2015.

Il s'agira dans un premier temps d'expliquer le principe général du partage des poids et une application optimale possible. Dans un second temps, nous verrons dans quelle mesure cette méthode s'applique à l'empilement annuel des bases de données du TCM : une première application se fera au niveau logement pour prendre en compte l'empilement d'échantillons d'enquêtes différentes et une deuxième application s'effectuera au niveau individu pour prendre en compte la multi-résidence. Enfin, nous terminerons par évoquer les autres stratégies possibles pour les utilisateurs de la base des TCM empilés.

9. Lorsque la définition de ménage se base sur la notion de budget commun, il est possible d'avoir plusieurs ménages dans un même logement. C'est le cas généralement de colocataires qui font chacun budget séparé.

Chapitre 1

La méthode généralisée du partage des poids

1.1 Le principe général

La méthode généralisée du partage des poids (MGPP) peut intervenir dans le cas d'un sondage indirect [1]. Plus précisément, dans le cas où l'unité d'observation, non seulement diffère de l'unité d'échantillonnage, mais encore peut être enquêtée suite au tirage de différentes unités d'échantillonnage. Cela peut arriver par exemple si l'on interroge un individu dans son logement, indépendamment de sa catégorie (principale / secondaire). Si cet individu possède une résidence principale et une résidence secondaire alors il pourra être enquêté, soit si sa résidence principale est tirée, soit si c'est sa résidence secondaire qui l'est.

Considérons une population Ω_A représentant la base de sondage dans laquelle sont les unités d'échantillonnage (logements par exemple) et une population Ω_B contenant les unités d'observation (individus par exemple). Un système de liens $L_{j,i}$ existe entre ces deux populations : pour tout j de Ω_A et pour tout i de Ω_B , $L_{j,i} > 0$ si j renvoie à i, et 0 sinon (dire que j renvoie à i signifie que si on échantillonne j alors on observera i). Dans la plupart des cas, le système de liens est défini de cette façon : $L_{j,i}$ vaut 1 s'il existe un lien entre i et j et 0 sinon. Nous verrons cette application dans la section 2.2 quand nous aborderons la pondération des individus prenant en compte la multi-résidence. La figure 1.1 permet de mieux visualiser ce système de liens entre les deux populations. Plusieurs flèches peuvent partir d'une unité j de Ω_A et plusieurs flèches peuvent arriver sur une unité i de Ω_B (cela se conçoit bien si j est un logement et i un individu : un logement peut en effet contenir plusieurs individus et un individu peut avoir plusieurs logements s'il est multi-résident).

Soit S_A un échantillon d'unités j de Ω_A. On obtient alors l'échantillon S_B d'unités i de Ω_B via le système de liens. Soit un total Y sur la population d'intérêt Ω_B, $Y = \sum_{i \in \Omega_B} Y_i$, que l'on va chercher à estimer à partir de l'échantillon S_B .
L'approche naturelle pour estimer sans biais Y consiste à prendre l'estimateur de Horvitz-Thompson et donc à calculer les probabilités d'inclusion des unités i. Mais il est très difficile d'obtenir ces probabilités en pratique car elles dépendent des probabilités de tirage des unités j de Ω_A qui renvoient à i et cela fait donc intervenir des probabilités d'inclusion d'ordre supérieur ou égal à deux [2].
C'est la méthode de partage des poids qui va permettre de contourner ce problème en conduisant à un estimateur sans biais de Y, qui n'est pas l'estimateur de Horvitz-Thompson. On peut montrer

1. Dans un sondage indirect, les unités d'observation ne sont pas tirées directement. On tire des unités d'échantillonnage et ce sont elles qui font le lien avec les unités d'observation. C'est ce qui est fait par exemple dans les enquêtes ménages où l'unité d'intérêt est un individu. En effet, comme on ne dispose que d'une base de logements, on va d'abord tirer un logement et c'est par le biais de ce logement que l'on pourra atteindre l'individu.

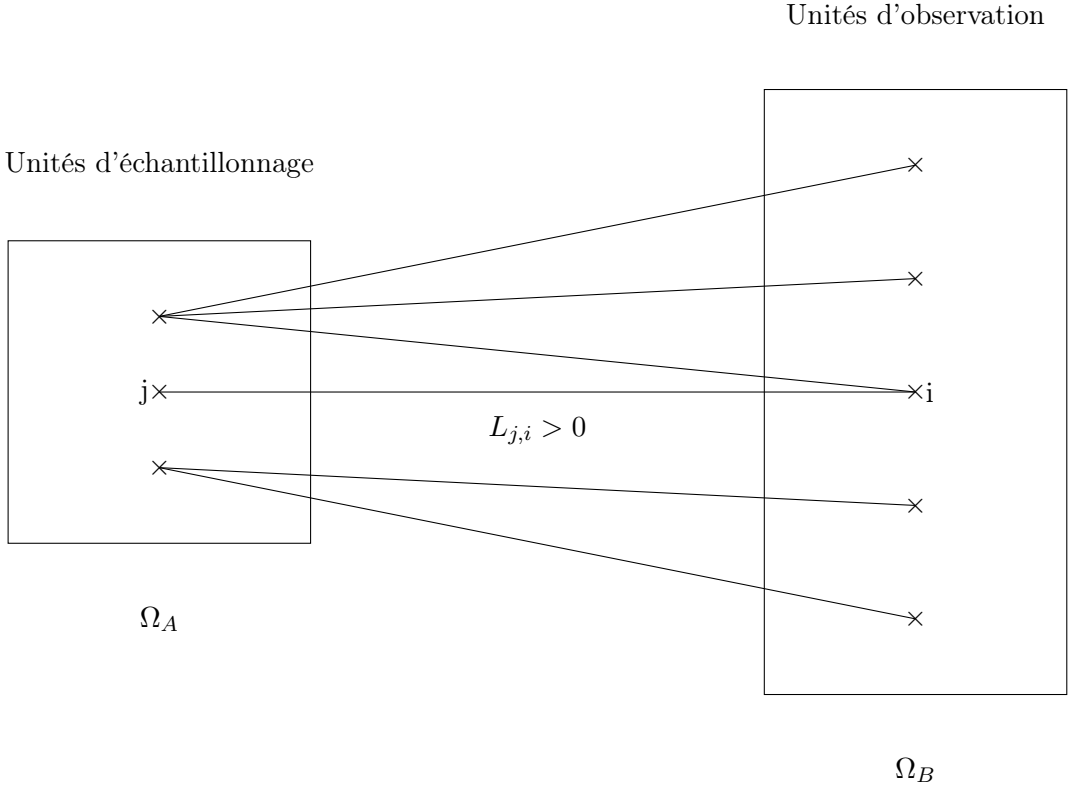


Figure 1.1 – Visualisation du système de liens entre unités d'échantillonnage et unités d'observation

que cet estimateur s'écrit

$$\hat{Y} = \sum_{i \in S_B} W_i.Y_i$$

où

$$W_i = \sum_{j \in S_A} \theta_j.\frac{L_{j,i}}{L_i} = \text{poids issu de la MGPP} \tag{1.1}$$

avec θ_j poids de sondage de j lié au tirage de S_A dans Ω_A et $Li = \sum_{j \in \Omega_A} L_{j,i}$ = nombre total de liens que l'unité i peut avoir avec les unités j de Ω_A. La démonstration figure en annexe A.
L'absence de biais de cet estimateur n'est garantie que s'il n'existe aucune unité i de Ω_B telle que $L_i = 0$. Si cela s'avérait être le cas, l'unité i serait inaccessible puisqu'aucune unité j échantillonnée dans Ω_A ne pourrait conduire à i. Il y aurait donc un défaut de couverture, se traduisant par un biais.

Dans l'annexe A, on montre que le total Y peut se réécrire sous la forme $Y = \sum_{j \in \Omega_A} Z_j$ où $Z_j = \sum_{i \in \Omega_B} \frac{L_{j,i}}{L_i}.Y_i$. L'estimateur issu de la MGPP peut alors s'écrire comme l'estimateur de Horvitz-Thompson d'un nouveau total Z [2] sur la population Ω_A , $\hat{Y} = \sum_{j \in S_A} \theta_j.Z_j$ et on se retrouve dans le cadre classique pour estimer la variance $\mathbb{V}\left[\hat{Y}\right] = \mathbb{V}\left[\sum_{j \in S_A} \theta_j.Z_j\right]$.

Ainsi, dans le cas où les unités d'intérêt peuvent être atteintes par le biais de différentes unités d'échantillonnage, il est possible de reconstruire un estimateur sans biais issu de la MGPP et pour lequel un calcul d'estimation de variance est réalisable.
Nous avons raisonné ici à partir des poids de tirage initiaux mais la MGPP est applicable également dans le cas où l'on dispose de poids issus de redressements ou de traitements de la non-réponse [3].

2. $Z = \sum_{j \in \Omega_A} Z_j$

C'est d'ailleurs le cas dans lequel nous serons pour les TCM empilés.

1.2 L'approche optimale de la MGPP

Très généralement, le système de liens utilisé dans la MGPP est celui qui vaut 1 si j renvoie à i et 0 sinon. Ce n'est pas le système optimal qui permet de minimiser la variance des estimateurs, mais ce dernier est difficile à obtenir en pratique, donc il est convenu de faire ce choix. Cependant, dans le cas particulier des bases de sondage multiples [3], nous allons voir qu'il est possible d'utiliser le système de liens optimal.

Considérons d'abord que l'échantillonnage s'effectue dans deux bases distinctes [4] pour obtenir à la fin un seul échantillon résultant. Dans la section précédente, le problème était qu'une unité d'observation pouvait être atteinte par le biais de plusieurs unités d'échantillonnage distinctes. Ici, le problème concerne directement les unités d'échantillonnage qui peuvent être tirées plusieurs fois au travers de différentes bases et nous ne sommes plus nécessairement dans le cadre d'un sondage indirect.

Soient Ω_1 et Ω_2 les deux populations, non nécessairement disjointes, associées à chacune des bases de sondage. Soient S1 (resp. S2) l'échantillon tiré dans Ω_1 (resp. Ω_2) et $L_{j1,i}$ (resp. $L^*_{j2,i}$) le système de liens associé. Les tirages de S_1 et S_2 sont indépendants. $\Omega = \Omega_1 \cup \Omega_2$ est la population résultante et donc le champ de l'enquête. $S = S_1 \cup S_2$ est l'échantillon résultant. On cherche à estimer le total $Y = \sum_{i\in\Omega} Y_i$.
Dans l'annexe B, on montre que

$$Y = \sum_{j1\in\Omega_1} Z_{j1} + \sum_{j2\in\Omega_2} Z^*_{j2} \tag{1.2}$$

où $Z_{j1} = \sum_{i\in\Omega} \frac{L_{j1,i}}{L_i}.Y_i$; $Z^*_{j2} = \sum_{i\in\Omega} \frac{L^*_{j2,i}}{L_i}.Y_i$ et

$$L_i = \sum_{j1\in\Omega_1} L_{j1,i} + \sum_{j2\in\Omega_2} L^*_{j2,i} \tag{1.3}$$

L_i est le nombre de liens qu'une unité d'échantillonnage i possède avec l'ensemble des bases.
Le schéma 1.2 permet de visualiser le système de liens dans le cas de deux bases de sondage distinctes.
L'estimateur de Y issu de la MGPP ($\hat{Y} = \sum_{i\in S} W_i.Y_i$) peut donc s'écrire aussi comme la somme des estimateurs de Horvitz-Thompson de deux nouveaux totaux Z_1 et Z^*_2 [5] sur les populations respectives Ω_1 et Ω_2, soit :

$$\hat{Y} = \sum_{j1\in S_1} \theta_{j1}.Z_{j1} + \sum_{j2\in S_2} \theta^*_{j2}.Z^*_{j2} \tag{1.4}$$

(θ_{j1} et θ^*_{j2} sont respectivement les poids de tirage des unités j1 et j2 dans les échantillons S_1 et S_2).
Cela permet de montrer (cf. annexe B) que le poids W_i de l'unité i dans l'échantillon S, issu de la MGPP, s'écrit :

$$W_i = \frac{1}{L_i}\left(\sum_{j1\in S_1} \theta_{j1}.L_{j1,i} + \sum_{j2\in S_2} \theta^*_{j2}.L^*_{j2,i}\right) \tag{1.5}$$

3. Par exemple, on est dans le cas de bases de sondage multiples lorsqu'une partie de l'échantillon final a été tirée dans l'annuaire téléphonique et l'autre partie dans la base des logements. Dans ce cas, tout ménage possédant une ligne téléphonique peut être tiré au travers des deux bases de sondage.

4. Ce raisonnement s'appliquerait de la même façon au cas de m bases de sondage distinctes.

5. $Z_1 = \sum_{j1\in\Omega_1} Z_{j1}$ et $Z^*_2 = \sum_{j2\in\Omega_2} Z^*_{j2}$

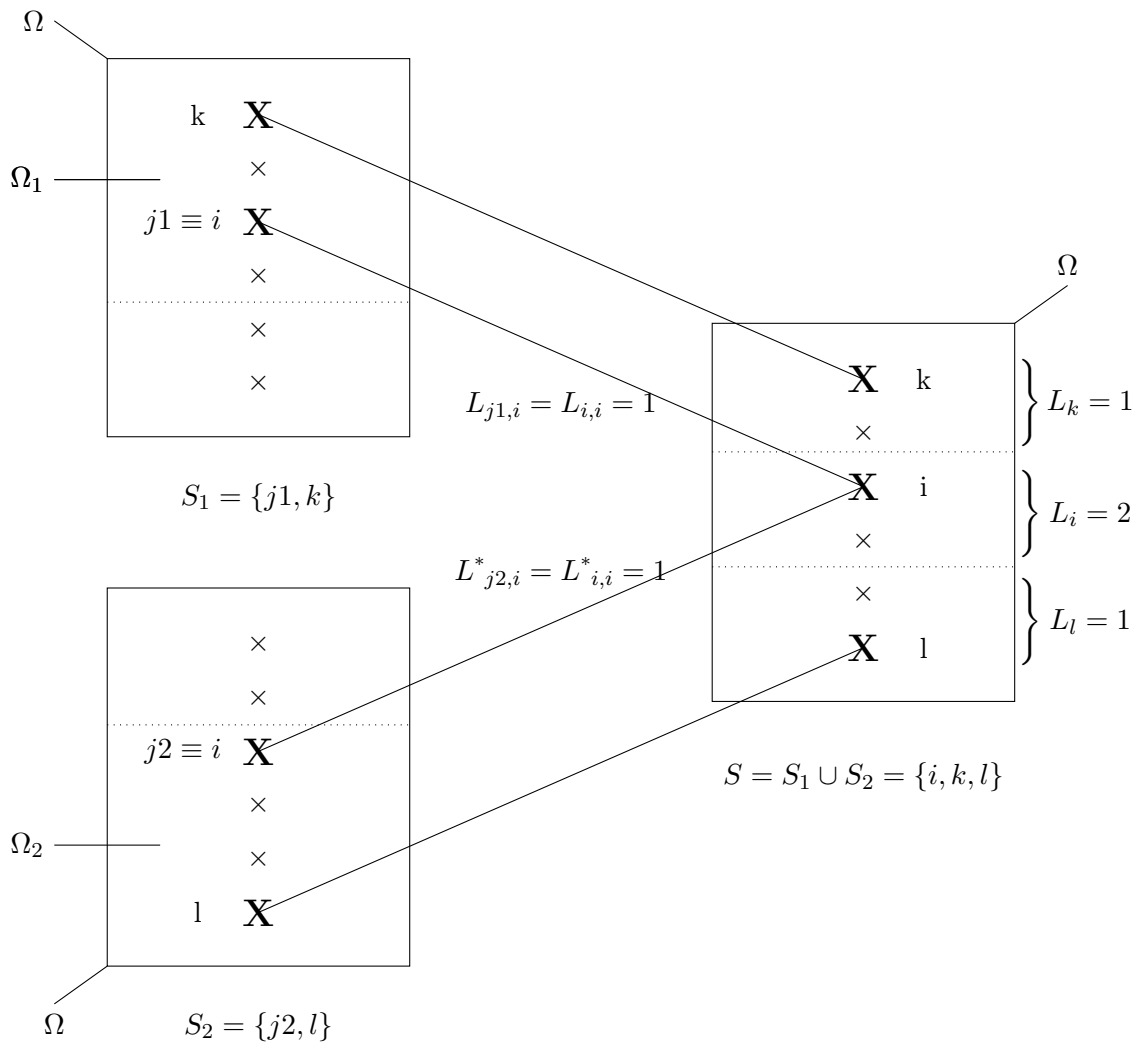


Figure 1.2 – Visualisation du système de liens dans le cas de deux bases de sondage distinctes

Ici, dans notre cas, comme le montre la figure 1.2, les unités d'observation i sont en fait les unités d'échantillonnage donc i renvoie à j1 ou j2 si, et seulement si i = j1 ou i = j2.
On peut donc réécrire la formule 1.5 sous cette forme :

$$W_i = \frac{1}{L_i}\Big(\sum_{j1\in S_1} \theta_{j1}.L_{j1,i}.\mathbb{1}_{j1=i} + \sum_{j2\in S_2} \theta^*_{j2}.L^*_{j2,i}.\mathbb{1}_{j2=i}\Big)$$

Ce qui permet d'obtenir l'expression finale :

$$W_i = \frac{1}{L_i}(\theta_i.L_{i,i}.\mathbb{1}_{i\in S_1} + \theta^*_i.L^*_{i,i}.\mathbb{1}_{i\in S_2}) \tag{1.6}$$

De même comme i et j sont de même nature, on peut réécrire la formule 1.3 sous la forme :

$$L_i = L_{i,i}.\mathbb{1}_{i\in\Omega_1} + L^*_{i,i}.\mathbb{1}_{i\in\Omega_2} \tag{1.7}$$

Considérons maintenant que les deux populations Ω_1 et Ω_2 sont identiques : $\Omega_1 = \Omega_2 = \Omega$. On tire deux échantillons S_1 et S_2 dans Ω et on considère l'échantillon S = $S_1 \cup S_2$. Soient n_1 (resp. n_2) la taille de l'échantillon S_1 (resp. S_2) et θ_i (resp. θ^*_i) le jeu de poids de tirage associé.
Le total Y est estimé sans biais par deux estimateurs de Horvitz-Thompson concurrents : $\hat{Y_1} = \sum_{i\in S_1} \theta_i.Y_i$ et $\hat{Y_2} = \sum_{i\in S_2} \theta^*_i.Y_i$. Nous raisonnons ici dans le cas de deux échantillons pour simplifier les notations mais ce raisonnement s'applique de la même façon dans le cas de m échantillons.

Cherchons l'estimateur optimal de Y sur l'échantillon S.
Il s'écrit sous forme d'une combinaison linéaire de $\hat{Y_1}$ et $\hat{Y_2}$: $\hat{Y}_{opti} = \alpha.\hat{Y_1} + \beta.\hat{Y_2}$, tel que $\alpha + \beta = 1$ pour obtenir un estimateur également sans biais.
On peut encore écrire

$$\hat{Y}_{opti} = \sum_{i \in S} \alpha.\theta_i.\mathbb{1}_{i \in S_1}.Y_i + \sum_{i \in S} \beta.{\theta^*}_i.\mathbb{1}_{i \in S_2}.Y_i = \sum_{i \in S} W_i^{opti}.Y_i$$

avec

$$W_i^{opti} = \alpha.\theta_i.\mathbb{1}_{i \in S_1} + \beta.{\theta^*}_i.\mathbb{1}_{i \in S_2} \tag{1.8}$$

Il s'agit donc de trouver le couple (α,β) qui va minimiser $\mathbb{V}\left[\hat{Y}_{opti}\right] = \alpha^2.\mathbb{V}\left[\hat{Y_1}\right] + \beta^2.\mathbb{V}\left[\hat{Y_2}\right]$ (les échantillons S_1 et S_2 sont indépendants). Après résolution du programme d'optimisation sous contrainte suivant (cf. résolution en annexe C),

$$\begin{cases} \min\limits_{(\alpha,\beta)} \mathbb{V}\left[\hat{Y}_{opti}\right] \\ \text{sous la contrainte } \alpha + \beta = 1 \end{cases}$$

on obtient :

$$\alpha = \frac{\mathbb{V}\left[\hat{Y_2}\right]}{\mathbb{V}\left[\hat{Y_1}\right] + \mathbb{V}\left[\hat{Y_2}\right]} \text{ et } \beta = \frac{\mathbb{V}\left[\hat{Y_1}\right]}{\mathbb{V}\left[\hat{Y_1}\right] + \mathbb{V}\left[\hat{Y_2}\right]}$$

Une approximation usuelle, vérifiée théoriquement dans le cas d'un sondage aléatoire simple ou stratifié et vérifiée empiriquement dans les autres cas, est de considérer que la variance est inversement proportionnelle à la taille de l'échantillon.
On obtient alors

$$\alpha = \frac{n_1}{n_1 + n_2} \text{ et } \beta = \frac{n_2}{n_1 + n_2}$$

À partir des expressions 1.6 et 1.8, on en déduit que :

$$W_i = W_i^{opti} \Leftrightarrow \frac{L_{i,i}}{L_i} = \alpha = \frac{n_1}{n_1 + n_2} \text{ et } \frac{{L^*}_{i,i}}{L_i} = \beta = \frac{n_2}{n_1 + n_2}$$

Ainsi, pour obtenir un système de liens optimal qui minimiserait la variance des estimateurs, on peut choisir dans chaque base un lien égal à la taille de l'échantillon tiré : $L_{i,i} = n_1, {L^*}_{i,i} = n_2$ et $L_i = n_1 + n_2 = n$ (comme cela a été dit précédemment dans cette section, si j est différent de i alors nécessairement, dans notre cas, j ne renvoie pas à i et donc $L_{j,i} = 0$).

Nous pouvons remarquer que si nous avions choisi au départ $L_{i,i} = {L^*}_{i,i} = 1$ (et donc L_i = nombre d'échantillons contenant i) alors il s'agirait du système de liens optimal uniquement dans le cas où les tailles d'échantillon seraient équilibrées (en effet, on aurait ici $\alpha = \frac{L_{i,i}}{L_i} = \frac{1}{2} = \beta$ or si $n_1 << n_2$ ou $n_2 << n_1$ alors $\frac{n_1}{n_1+n_2} \neq \frac{1}{2}$ (donc $\neq \alpha$) et $\frac{n_2}{n_1+n_2} \neq \frac{1}{2}$ (donc $\neq \beta$)).

Pour récapituler, nous distinguons donc deux méthodes. La première s'applique dans le cas général d'un sondage indirect (et donc aussi dans le cas de bases de sondage multiples). L'unité d'observation peut être atteinte à travers plusieurs unités d'échantillonnage. Mais on ne connaît pas le système de liens optimal et on choisit par défaut le système « 0-1 » même s'il ne minimise pas la variance des estimateurs. La seconde s'applique uniquement dans le cas de bases de sondage multiples, plus précisément, dans le cas où l'unité d'observation ou d'échantillonnage peut être tirée dans plusieurs bases ou plusieurs échantillons d'une même base. C'est la seule méthode qui permette d'utiliser le système de liens optimal. Nous allons maintenant expliquer dans quelles mesures ces deux méthodes s'appliquent dans le cas des TCM empilés.

Chapitre 2

Quels systèmes de pondération pour les tables des TCM empilés ?

À l'aval de chaque enquête embarquant le TCM, sont produites trois tables TCM correspondant aux trois niveaux logement, ménage et individu.
Comme indiqué précédemment, nous procédons à deux types d'empilement des TCM. Il y a d'abord l'empilement annuel qui consiste à empiler, pour chaque niveau, les tables TCM issues des enquêtes d'une même année. On obtient ainsi trois tables TCM annuelles pour chaque année entre 2006 à 2015. Ensuite, nous procédons à l'empilement périodique qui consiste à empiler, pour chaque niveau, les tables TCM annuelles de 2006 à 2015.

Il va donc s'agir ici de déterminer, pour chaque empilement (annuel et périodique) et pour chaque niveau, les systèmes de poids finaux.
Les poids de départ pour les tables TCM issues de chaque enquête sont les poids finaux de niveau logement/ménage [1] fournis par chaque responsable d'enquête et utilisés pour leurs exploitations usuelles. Ils sont déjà corrigés de la non-réponse et calés. Nous aurions également pu partir des poids de tirage initiaux mais cela aurait impliqué de refaire tout le traitement de la non-réponse alors qu'il a déjà été fait pour chaque enquête. Par ailleurs, la correction de la non-réponse à l'issue d'une enquête tient compte de ses particularités, ce que nous ne pourrions pas faire à notre niveau. Les modèles de non-réponse risqueraient donc d'être moins bons.

Nous allons ici commencer par appliquer l'approche optimale du partage des poids dans le cas de l'empilement annuel des tables TCM de niveau logement. Ensuite, nous expliquerons l'application du principe général du partage des poids dans le cas de l'empilement annuel des tables TCM de niveau individu. Et enfin, nous verrons l'empilement des tables TCM annuelles sur la période 2006-2015.

2.1 Détermination des poids logements annuels

2.1.1 Une application optimale de la méthode du partage des poids

Dans un premier temps, il s'agit d'empiler les tables TCM niveau logement des enquêtes d'une même année. Pour chaque enquête, on dispose d'un jeu de poids corrigés de la non-réponse et calés. Il sera noté $pondlog_k^{enq}$ (enq fait référence à l'enquête concernée et k au logement).
On se retrouve dans le cadre expliqué à la section 1.2. En effet, au départ nous disposons d'un échantillon S_{enq}^{log} pour chaque enquête. Ces échantillons ne sont pas nécessairement tirés dans la

1. Il n'existe pas de marges au niveau ménage donc les calages s'effectuent au niveau logement et lorsque cela est nécessaire, on attribue à chaque ménage, le poids de son logement.

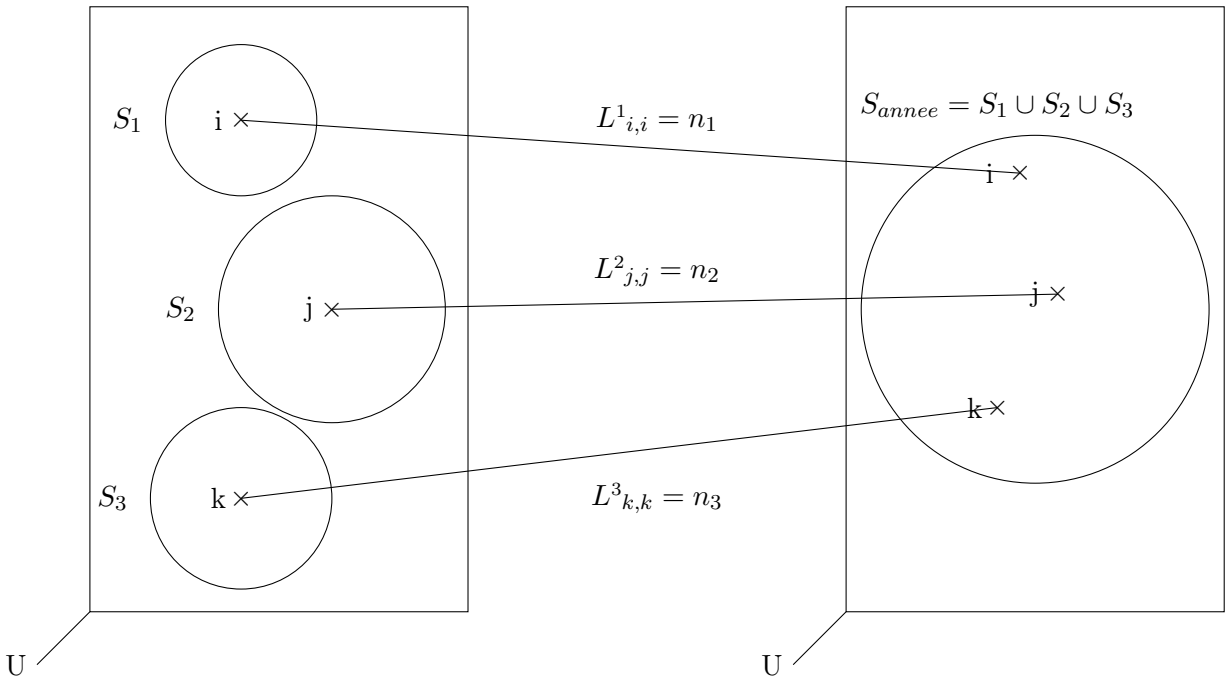


Figure 2.1 – Aproche optimale de la MGPP dans le cas de l'empilement de trois enquêtes

même base : il peut s'agir de l'échantillon-maître du recensement de la population ou des fichiers de la Taxe d'Habitation (TH). Mais chacun extrapole bien sur la même population U : celle des logements ordinaires en résidence principale (c'est en effet une condition nécessaire à l'empilement [2]). En empilant toutes ces enquêtes d'une même année, on crée un nouvel échantillon S^{log}_{annee} qui est la réunion des échantillons S^{log}_{enq} et pour lequel on veut déterminer un jeu de poids extrapolant sur U. On peut donc appliquer la méthode optimale du partage des poids et choisir comme système de liens $L^{enq}_{k,k} = n_{enq}$ = taille de l'échantillon de répondants [3] à l'enquête *enq* (on a également $L_k = n =$ taille totale de l'échantillon de répondants S^{log}_{annee} issu de l'empilement $= \sum_{enq} n_{enq}$). La figure 2.1 résume la situation dans le cas de l'empilement de trois enquêtes 1, 2 et 3 (on simplifie les notations en enlevant la précision *log* pour le niveau logement).

Les tirages d'échantillons sont organisés de telle sorte que si un logement est tiré pour une enquête une année donnée alors il ne pourra pas l'être pour une autre enquête cette même année. L'expression du poids issu de la MGPP ne fait donc intervenir qu'un seul terme correspondant à une enquête. Le nouveau jeu de poids $pondlog\_PP\_annee$ [4] associé à l'échantillon empilé S^{log}_{annee} est donc défini de cette façon :

$$\boxed{\forall k \in S^{log}_{annee} : \qquad pondlog\_PP\_annee_k = \frac{n_{enq}}{n} \times pondlog^{enq}_k}$$

où *enq* correspond à l'enquête pour laquelle le logement k a été interrogé.

2.1.2 Calage niveau logement

À l'issue du partage des poids précédent, un calage supplémentaire est effectué sur le jeu de poids $pondlog\_PP\_annee$. Certes, un calage a déjà été effectué en amont pour chaque enquête. Toutefois, les responsables d'enquête n'ont pas forcément utilisé les mêmes variables de calage, ce qui ne

2. Dans le TCM, les responsables d'enquête ont la possibilité d'interroger plus largement les ménages dans leur résidence habituelle, même si elle n'est pas principale. Cependant, jusqu'à présent le champ couvert est toujours celui des logements ordinaires qui sont la résidence principale pour au moins une personne d'un ménage.

3. Ce n'est pas la taille de l'échantillon initial puisque la non-réponse a déjà été traitée. Dans la table, ne figurent donc que les logements répondants et pondérés à l'issue des traitements de l'enquête.

4. PP corrrespond à Partage des Poids.

garantit donc pas nécessairement une cohérence avec les mêmes totaux et effectifs. Par ailleurs, cela permettra d'utiliser les variables de calage qui semblent les plus pertinentes pour l'exploitation qui sera faite de la base des TCM empilés.
En revanche, il n'est pas nécessaire de refaire un traitement de la non-réponse. En effet, celui-ci ayant déjà été réalisé en amont, asymptotiquement, il n'y a plus de biais de non-réponse.
À l'issue de ce calage sur les poids issus de la MGPP, le nouveau jeu de poids obtenu aboutira donc encore à des estimateurs asymptotiquement sans biais.

Les différents projets d'exploitation de l'Ined feront intervenir les caractéristiques sociodémographiques des individus et de leur ménage. Nous utiliserons des variables de calage assez classiques recommandées dans différentes notes de l'Insee et potentiellement liées avec les thèmes d'intérêt des utilisateurs de la base des TCM empilés : zone géographique dans laquelle se trouve le logement ; type de ménage ; tranche d'âge, catégorie socioprofessionnelle et diplôme de la personne de référence du logement. Il s'agit ici de variables catégorielles, dont les marges de référence (ou effectifs) fournies par la division Sondages sont issues de l'Enquête Emploi en Continu (EEC). Le nombre total de logements en résidence principale sera également utilisé pour le calage mais il provient du bilan annuel du logement 2017, qui fournit la meilleure estimation possible pour les années 1999 à 2017.

À titre d'exemple, figurent ici les résultats du calage effectué sur les poids $pondlog\_PP\_2011_k$ pour l'année 2011.
La table annuelle *TCM_LOG_2011* empile les enquêtes CVS (Cadre de Vie et Sécurité), BdF (Budget des Familles) et SRCV (Statistiques sur les Ressources et Conditions de Vie des ménages). Elle contient 38 521 logements répondants.
Le calage a été réalisé à l'aide de la macro SAS calmar2 [4]. Avant le calage, l'échantillon est peu déséquilibré par rapport à la population pour les variables auxiliaires utilisées. Cela signifie que les estimations des marges, obtenues avec le jeu de poids à corriger, sont proches des marges de référence issues de l'EEC. Il ne sera donc pas nécessaire de distordre énormément les poids de départ pour estimer parfaitement ces marges de référence. On s'attend ainsi à ce que les rapports de poids (poids final/poids initial à corriger) aient un « bon comportement »(faible dispersion, faible étendue et distribution gaussienne centrée en 1).
Les quatres méthodes classiques [5] de calage ont été appliquées. Elles donnent sensiblement les mêmes résultats. C'est la méthode raking-ratio qui a finalement été choisie car c'est celle utilisée habituellement dans le cas de variables auxiliaires qualitatives et l'étendue des rapports de poids étant faible, il n'a pas semblé nécessaire de choisir une méthode bornée.
Voici les résultats obtenus pour la distribution des rapports de poids avec la méthode raking-ratio :

- étendue de 0,92 avec des rapports de poids compris entre 0,65 et 1,57 ;
- écart-type de la distribution égal à 0,13 ;
- distribution gaussienne centrée en 1 (cf. figure 2.2 qui présente l'allure de la distribution obtenue pour les rapports de poids).

Remarque : la méthode choisie n'est pas nécessairement la même pour toutes les années puisque le choix dépend de l'allure de la distribution des rapports de poids.

5. Il s'agit des méthodes linéaire, linéaire tronquée, raking-ratio et logit (raking-ratio tronquée).

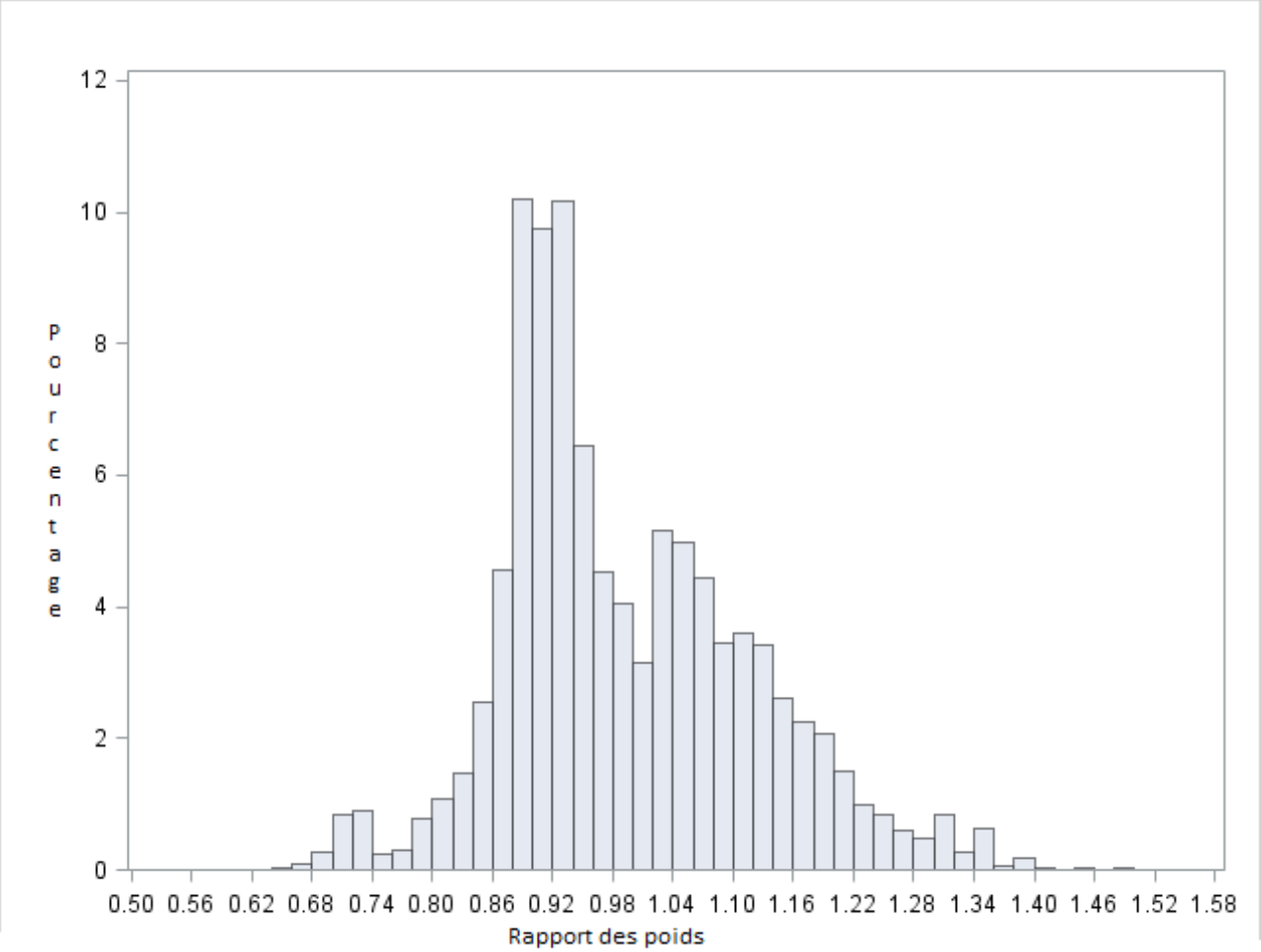


Figure 2.2 – Histogramme des rapports de poids issus de la méthode raking-ratio
Source : Table des TCM empilés pour l'année 2011 au niveau logement (enquêtes CVS, SRCV et BDF)
Effectifs : 38 521 logements ordinaires en résidence principale

Et enfin, une fois le calage terminé, nous obtenons les poids logements annuels finaux $pondlog\_final\_annee_k$ dans chaque table TCM_LOG_annee.
Pour les poids ménages finaux des tables TCM_MEN_annee, on affecte simplement à chaque ménage le poids de son logement, qu'il y ait un seul ou plusieurs ménages dans le logement. En effet, s'il n'y a qu'un seul ménage dans le logement alors logement et ménage sont équivalents donc ils ont le même poids. Et s'il y a plusieurs ménages dans le logement, même s'ils ne sont pas tous dans le champ de l'enquête, ils seront nécessairement tous interrogés dans le TCM. On affecte donc le poids du logement à tous les ménages.
Il s'agit ensuite de déterminer le poids des individus de chaque ménage dans les tables TCM_IND_annee.

2.2 Détermination des poids individus annuels

Tous les individus d'un ménage étant systématiquement interrogés pour le TCM, il s'agit d'un tirage par grappe[6]. Dans un premier temps, on attribue donc à chaque individu du ménage le poids de son ménage.
Cependant, pour chaque enquête à empiler sur la période 2006-2015, 4 à 7 % des répondants sont multi-résidents. Dès lors, attribuer à chaque individu, le poids de son ménage, peut induire un

6. Cas particulier d'un sondage indirect où toutes les unités secondaires d'une même unité primaire sont interrogées.

biais. En effet, il serait plus rigoureux de tenir compte du fait qu'une personne ayant plusieurs logements dans le champ de l'enquête a une probabilité d'inclusion proportionnelle au nombre de ses logements. Par ailleurs, dans l'article présenté aux JMS en 2012 [1], Laurent Toulemon et Thomas Denoyelle montrent que lorsque la multi-résidence est ignorée dans les pondérations, elle peut perturber l'estimation de certains indicateurs décrivant les situations familiales, comme le nombre de familles monoparentales par exemple.
Il a donc été décidé de tenir compte de la multi-résidence dans les pondérations individuelles.

La méthode de partage des poids prend alors tout son sens puisque tenir compte de la multi-résidence des individus revient à se placer dans le cadre général expliqué à la section 1.1. En effet, nous distinguons les logements (ou ménages), unités d'échantillonnage, des individus, unités d'observation, et, par définition, un individu multi-résident peut être atteint par le biais de plusieurs unités d'échantillonnage (pourvu que ses autres logements soient ordinaires en résidence principale[7]). Un système de liens $L_{j,i}$ peut donc être défini entre les logements j et les individus i. Il s'agira du système de liens classique valant 1 si l'individu i peut être interrogé via le logement j et 0 sinon. Le nombre total de liens qu'un individu i possède avec la population des logements est donc le nombre de logements ordinaires dans lesquels il vit habituellement (on le notera Nb_log_i[8]).

Soient S_{log} l'échantillon de logements de la table *TCM_LOG_annee* et S_{ind} l'échantillon d'individus de la table *TCM_IND_annee*. En appliquant la formule 1.1 au cas des TCM empilés, le poids d'un individu i du logement j issu de la MGPP est :

$$pondind\_PP\_theorique\_annee_{i,j} = \sum_{l \in S_{log}} \frac{pondlog\_final\_annee_l}{Nb\_log_i}$$

Mais en pratique, si un individu i est présent plusieurs fois dans la table *TCM_IND_annee* par le biais de ses différents logements, nous ne pouvons pas l'identifier en tant que doublon car il aura des identifiants différents. Cela signifie que les poids de ses autres logements qui seraient éventuellement dans l'échantillon S_{log} ne sont pas connus à travers une seule observation. Dès lors, on redéfinit le poids de l'individu i du logement j de cette façon :

$$pondind\_PP\_annee_{i,j} = \frac{pondlog\_final\_annee_j}{Nb\_log_i}$$

Mais finalement nous sommes toujours bien dans le cadre de la MGPP. En effet, admettons que l'individu i ait ses deux logements j et k tirés dans l'échantillon S_{log}, alors il aura deux poids associés à ses deux observations dans la table : $pondind\_PP\_annee_{i,j} = \frac{pondlog\_final\_annee_j}{Nb\_log_i}$ et $pondind\_PP\_annee_{i,k} = \frac{pondlog\_final\_annee_k}{Nb\_log_i}$.
Puis, on remarque qu'en sommant ces deux poids, on retrouve bien le poids issu de la MGPP pour cet individu i.

Ensuite, une fois que le jeu de poids prenant en compte la multi-résidence est déterminé, on effectue un calage supplémentaire au niveau individuel. Les variables auxiliaires utilisées seront le sexe et la tranche d'âge car les phénomènes d'intérêt des utilisateurs de la base des TCM empilés sont susceptibles de varier fortement avec l'âge (le phénomène de multi-résidence notamment [1]). En particulier, nous utilisons comme variable de calage, le croisement sexe × tranche d'âge. Nous

7. Les questions du TCM ne permettent pas de savoir si l'autre logement d'un individu est une résidence principale pour un ménage. Ainsi, dès lors que l'autre logement est ordinaire (cette information en revanche peut être recueillie via le TCM), on considérera qu'il est potentiellement une résidence principale et donc qu'il appartient bien à la population de la base de sondage.

8. Cela correspond à la variable NAUTLOG_IND dans les tables *TCM_IND*.

disposons, ici encore, des marges de référence issues de l'EEC pour les années 2006 à 2015.
À l'issue de ce calage effectué dans chaque table *TCM_IND_annee*, nous obtenons les poids individus annuels finaux que l'on notera $pondind\_final\_annee_{i,j}$ pour un individu i du logement j.

2.3 Empilement des tables TCM annuelles sur la période 2006 à 2015

Après avoir empilé les enquêtes pour chaque année et déterminé les poids logements et individus annuels, il faut empiler les tables TCM annuelles sur la période 2006-2015 pour chaque niveau (logement, ménage et individu). On peut empiler sur la période car toutes les années représentent le même champ, celui des logements ordinaires en résidence principale.
On notera *TCM_LOG_06-15*, *TCM_MEN_06-15* et *TCM_IND_06-15* les tables TCM périodiques empilant les dix tables TCM annuelles.
Les exploitations sur la base des TCM empilés ont vocation à étudier les phénomènes en moyenne sur l'ensemble de la période puisqu'on considère qu'il n'y a pas de tendance temporelle [9]. Une estimation périodique [10] sera donc simplement la moyenne des dix estimations annuelles.

Soit Y un total niveau logement que l'on cherche à estimer à partir de la base des TCM empilés périodique. Pour chaque année, l'estimation correspondante est

$$\hat{Y}_{annee} = \sum_{k \in S^{log}_{annee}} pondlog\_final\_annee_k . Y_k$$

où S^{log}_{annee} est l'échantillon de répondants associé à la table *TCM_LOG_annee*.
L'estimation moyenne obtenue avec l'échantillon périodique S^{log}_{06-15} de la table *TCM_LOG_06-15* est :

$$\hat{Y}_{06-15} = \frac{1}{10}\sum_{j=06}^{15} \hat{Y}_j = \frac{1}{10}\sum_{j=06}^{15}\sum_{k \in S_j} pondlog\_final\_annee^j{}_k . Y_k = \sum_{k \in S_{06-15}} \frac{1}{10} pondlog\_final\_annee_k . Y_k$$

Cela s'applique de la même façon pour un total niveau individu.
On en déduit alors facilement la pondération d'un logement k et d'un individu i du logement k, interrogés au cours de l'année *annee*, dans les tables TCM périodiques :

$$\boxed{pondlog\_06-15_k = \frac{1}{10} pondlog\_final\_annee_k}$$

et

$$\boxed{pondind\_06-15_{i,k} = \frac{1}{10} pondind\_final\_annee_{i,k}}$$

C'est ici la stratégie de pondération qui a été choisie par la division RTI (Recueil et Traitement de l'Information) du DMS (Département des Méthodes Statistiques), en collaboration avec la division Sondages, mais nous allons évoquer, dans le chapitre qui suit, les autres pistes qui auraient pu être envisagées, et qui pourront l'être, par les utilisateurs de la base des TCM empilés qui le souhaiteraient.

9. S'il y en a une, on ne travaillera que sur les tables annuelles.
10. Estimation issue de l'empilement des tables TCM de 2006 à 2015.

Chapitre 3

Limites de la stratégie choisie et autres pistes envisageables pour la pondération des TCM empilés

3.1 Solution alternative au traitement de la multi-résidence pour la pondération individuelle

En appliquant la MGPP pour traiter la multi-résidence, tous les individus d'un même ménage n'auront plus nécessairement le même poids. Cela risque de poser un problème de cohérence pour les études menées sur des entités pluri-individuelles (ménages, logements, couples, etc.). Par exemple, si les individus d'un même ménage n'ont pas le même poids, le nombre de femmes en couple hétérosexuel [1], d'hommes en couple hétérosexuel et de couples hétérosexuels ne seront pas forcément identiques.

Afin d'assurer cette cohérence au niveau pluri-individuel, se pose alors la question d'une solution alternative à la MGPP, qui permettrait à la fois de traiter la multi-résidence et d'attribuer le même poids aux individus d'un même ménage.
L'autre solution envisagée serait donc la suivante :

1. tirage aléatoire de la moitié des multi-résidents
2. affectation du poids du logement pour les multi-résidents tirés et pour les individus n'ayant pas d'autres logements
3. affectation d'un poids nul pour l'autre moitié des multi-résidents

Montrons que cette solution alternative conduit bien à un estimateur sans biais.

Soit S_{log} un échantillon de logements, avec pour jeu de poids associé, w_l, conduisant à des estimations non biaisées.
On effectue un tirage par grappe pour sélectionner les individus : l'échantillon d'individus, S_{ind}, est constitué de tous les individus i de chaque logement l et chacun a le poids de son logement. On a donc pour i $\in$ l, $w_{i,l} = w_l$.
Ensuite, on effectue un partage des poids sur l'échantillon S_{ind}. On obtient ${w_i}^{\text{mgpp}} = \frac{w_{i,l}}{Nb\_log_i}$ (en faisant l'hypothèse qu'aucun individu multi-résident n'a un autre de ses logements dans S_{log}). Notons ${S_{ind}}^{\text{mgpp}}$ l'échantillon associé. On sait que dans le cadre de la MGPP, cet échantillon conduit à des estimations sans biais.

1. Sous-entendu, couple vivant dans le même logement.

Enfin, on effectue dans ${S_{ind}}^{\text{mgpp}}$, un tirage d'échantillon à probabilités inégales $\frac{1}{Nb\_log_i}$. Dans le cadre de l'échantillonnage à deux phases, on sait que les nouveaux poids ${w^*}_i = {w_i}^{\text{mgpp}} \times Nb\_log_i$ des individus de l'échantilon tiré ${S^*}_{ind}$ conduisent encore à des estimations sans biais. Or on remarque que ${w_i}^{\text{mgpp}} \times Nb\_log_i = w_{i,l} = w_l$. On retrouve donc ${w^*}_i = w_l$ et la solution alternative conduirait aussi à des estimateurs non biaisés.

Cependant, même si cette méthode permet d'obtenir des estimateurs sans biais [2], elle est moins efficace que la méthode de partage des poids. En effet, dans le cadre d'un sondage à deux phases, il existe un terme de variance additionnel, dû à la source d'aléa supplémentaire causé par le deuxième tirage. Donc ici le tirage des multi-résidents va nécessairement introduire un terme de variance supplémentaire par rapport au cadre de la MGPP.
Par ailleurs, en tirant la moitié des individus multi-résidents, on fait l'hypothèse qu'ils ont systématiquement un seul autre logement. C'est le cas pour la majeure partie en effet (96.2 % des multi-résidents dans la base des TCM empilés de 2006 à 2015), mais en théorie, il faudrait procéder à un tirage à probabilités inégales, égales à l'inverse du nombre d'autres logements. Néanmoins, cela est plus compliqué à mettre en oeuvre et ça ne permet plus d'avoir un poids identique pour tous les individus d'un même logement.
De plus, mettre des poids nuls à certains individus peut perturber la cohérence entre la taille réelle du ménage et les observations sur lesquelles on travaille *in fine* dans la table individu, puisqu'elles seraient restreintes à celles ayant un poids non nul. Cela signifie également qu'on pourrait être amené à mettre un poids nul à la personne de référence du ménage, ce qui peut poser problème pour le calage après le tirage.

Ainsi, la production d'un autre jeu de poids individus, mobilisant un tirage aléatoire plutôt qu'un partage des poids n'a pas été réalisée et est laissée à la charge des utilisateurs s'ils le souhaitent.

3.2 Existence de doublons dans la table TCM périodique

Dans la table TCM périodique, toutes les observations ne sont pas distinctes car l'enquête SRCV est une enquête longitudinale. On risque donc de sous-estimer la variance du fait de la présence de doublons. En effet, faisons l'hypothèse usuelle que la variance varie en 1/taille d'échantillon. Arrondissons la taille de l'échantillon, de la table TCM périodique de niveau individu, à 1 million d'unités (987 163 observations exactement), dont 150 000 doublons liés à SRCV. Une estimation de variance, en considérant les autres unités comme distinctes [3], conduit alors à une sous-estimation de l'ordre de 15 %, ce qui n'est tout de même pas négligeable. Cependant, les caractéristiques des individus étant amenées à changer au fil du temps, deux observations correspondant au même individu ne seront pas nécessairement identiques du point de vue de l'information qu'elles contiennent. Ne pas traiter le cas des doublons de l'enquête SRCV semble donc acceptable à partir du moment où les utilisateurs de la base sont conscients de cette limite.

Par ailleurs, même si le tirage des échantillons des enquêtes ménages est organisé de telle sorte qu'un même logement ne puisse pas être tiré plusieurs fois sur un laps de temps rapproché, la période d'empilement étant ici de 10 années, le risque d'avoir ce type de doublons n'est pas nul. Ce dernier est tout de même faible, mais si la livraison des TCM empilés perdure dans le temps, il serait préférable de faire dorénavant des empilements sur cinq années. Ce d'autant plus que l'hypothèse

2. À condition de ne pas supprimer les multi-résidents non tirés mais leur affecter simplement un poids nul.
3. On va voir dans le paragraphe suivant que cela n'est pas assuré au vu de la longueur de la période d'empilement. Néanmoins, la probabilité qu'un même logement soit enquêté deux fois sur la période est tout de même faible.

d'invariance temporelle faite pour la détermination des poids périodiques (cf. section 2.3) serait plus vraisemblable sur cinq années que sur dix.

3.3 Utiliser d'autres marges de calage ?

L'utilisation des marges de l'enquête EEC pour les calages des TCM empilés peut être discutée en raison, notamment, d'une non prise en compte des configurations familiales particulières dans la variable de calage type de ménage. En effet, les principaux centres d'intérêt pour les études à partir des TCM empilés étant les structures familiales complexes, il semble légitime de chercher à utiliser des marges issues d'une enquête captant au mieux ces situations familiales, ce qui n'est pas le cas de l'enquête EEC entre 2006 et 2015.
L'enquête sur la Famille et les Logements 2011 (EFL 2011) pourrait être intéressante car elle capte précisément les structures familiales complexes mais cela poserait certainement un problème pour la fraîcheur de l'information puisqu'il n'y a eu qu'une seule enquête sur la période 2006-2015.
On pourrait également penser aux marges du recensement, mais de même, les structures familiales complexes ne sont précisément captées que depuis la refonte de la Feuille Logement, en place depuis le RP 2018.
Il est donc laissé à la charge des utilisateurs de procéder à un calage à partir d'autres marges que celles de l'EEC s'ils le souhaitent. Cela est possible puisque les poids intermédiaires issus du partage des poids ont été laissés dans les bases.

3.4 Un effet enquête persistant

Malgré le partage des poids prenant en compte la taillle d'échantillon de chaque enquête et malgré un calage supplémentaire qui permet d'utiliser des variables auxiliaires communes pour chaque enquête, il n'est pas possible de négliger l'effet enquête. En effet, il existe des facteurs externes que l'on ne peut pas maitriser.
Il s'agit d'abord du thème de l'enquête qui peut expliquer une différence dans la façon de recueillir certaines informations du TCM. Par exemple, concernant la multi-résidence, on peut s'attendre à ce que le bloc Lieux de vie du TCM soit mieux renseigné dans les enquêtes Transport et Logement que dans l'enquête Handicap Santé ou encore CVS puisqu'on conçoit bien que les résidences multiples sont plus proches du thème des premières enquêtes que des dernières.
D'autres sources de différences peuvent être citées : grève durant une période donnée qui perturbe la collecte, dates de collecte parfois moins propices pour contacter les ménages (périodes de vacances scolaires par exemple), événement politique qui fait qu'une enquête passe plus ou moins bien auprès des ménages, etc.
Enfin, des problèmes ponctuels dans le data-model[4] du TCM peuvent aussi être à l'origine de différences entre enquête. Par exemple, pour les enquêtes CVS 2010 et 2011, un problème est survenu dans le recueil des liens père-mère car les questions sur la situation des parents (vit dans le logement/vit ailleurs/est décédé/parent inconnu) n'étaient plus posées.
Toutes ces différences entre enquêtes peuvent être une source de variance supplémentaire dans les estimations mais celle-ci ne pourra pas être captée, même avec la stratégie de pondération présentée dans le chapitre 2.
Toutefois, pour permettre aux utilisateurs de cerner ces différences entre enquêtes, une variable « libellé de l'enquête » est ajoutée dans la table des TCM empilés (variable LIB_ENQ) et les spécificités pour chacune d'elles sont recensées dans la deuxième partie de ce document.

4. Le data-model est la version électronique du questionnaire sur le poste Capi des enquêteurs.

Conclusion

Nous avons ainsi présenté la stratégie de pondération adoptée pour la base des TCM empilés. Elle peut se résumer de cette façon :

I. Empilement des enquêtes d'une même année

   1. Table TCM annuelle niveau logement :

      i. Poids de départ = poids logements fournis par chaque responsable d'enquête, corrigés de la non-réponse et calés

      ii. Application de la méthode optimale du partage des poids pour obtenir un jeu de poids logements intermédiaires

      iii. Calage simple niveau logement pour obtenir le jeu de poids logements finaux

   2. Table TCM annuelle niveau ménage : affectation du poids logement final à chaque ménage du logement : obtention des poids ménages finaux

   3. Table TCM annuelle niveau individu :

      i. Affectation du poids ménage final à chaque individu du ménage

      ii. Application de la méthode du partage des poids pour prendre en compte la multi-résidence : obtention d'un jeu de poids individus intermédiaires

      iii. Calage simple niveau individu pour obtenir le jeu de poids individus finaux

II. Empilement des tables TCM annuelles : trois tables TCM périodiques – logement, ménage, individu – dont les jeux de poids finaux seront ceux des tables TCM annuelles – logement, ménage, individu - rapportés au nombre d'années empilés.

Il s'agit ici de la stratégie choisie par la division RTI mais tous les jeux de poids intermédiaires ont été laissés dans les tables, de manière à laisser à chaque utilisateur la liberté de produire des jeux de poids définitifs se basant sur une autre stratégie. Il pourra s'agir par exemple d'un tirage aléatoire des multi-résidents pour assurer des poids identiques entre tous les individus d'un même ménage [5] ou encore d'utiliser d'autres marges de calage.

5. Cette solution impliquerait néanmoins de supprimer du fichier un multi-résident sur deux.

Deuxième partie

Documentation afférente à la base des TCM empilés

Introduction

La partie méthodologie étant maintenant présentée, cette deuxième partie a vocation à présenter toute la documentation utile aux différents utilisateurs de la base.

Il s'agit dans un premier temps de lister toutes les enquêtes empilées avec les différents effectifs.
Ensuite, figure le dictionnaire des variables restreint aux principales variables d'intérêt. Pour le moment, toutes les variables de collecte, pas nécessairement utiles pour les exploitations ont été gardées. Il faudra encore faire un tri pour alléger les tables et en simplifier la compréhension. Une fois ce tri effectué, un nouveau dictionnaire des variables, complet, sera diffusé.
Enfin, sont listées les spécificités ou les problèmes observés sur certaines enquêtes.

Deux types de tables ont été mis à disposition : les tables finales avec uniquement les observations pondérées et les tables brutes comprenant également les logements déchets.
La documentation présentée ici ne concerne que les tables finales pondérées qui seront normalement les seules à être exploitées.
Dans les tables brutes, en plus des déchets non pondérés, figurent les observations associées aux enquêtes AES 2012 (Adult Education Survey) et CdT 2012 (Conditions de Travail). Celles-ci n'ont pu être intégrées aux tables finales pondérées car ces deux enquêtes ne possèdent pas de pondération niveau ménage mais uniquement une pondération au niveau individu et cela aurait été très lourd de recalculer des poids de niveau logement au vu de la complexité de l'échantillonnage. Par ailleurs, dans les tables brutes figurent aussi les fiches-adresses des DOM. Ces derniers n'ont pas été pris en compte dans les tables finales car les poids de niveau logement fournis par les concepteurs d'enquête pour les fiches-adresses de la France métropolitaine n'apportent une représentativité qu'à son échelle et non à celle de la France entière. Ainsi, il n'aurait pas été cohérent de combiner les poids des DOM avec le reste.

Chapitre 4

Liste des enquêtes constituant la base des TCM empilés

Dans le tableau 4, figure la liste de toutes les enquêtes empilées, ainsi que les effectifs des échantillons de répondants aux niveaux logement, ménage et individu [1].

Enquête	**Effectif des logements**	**Effectif des ménages**	**Effectif des individus**
CVS 2006	7 251	7 295	17 812
Logement 2006	36 955	37 298	93 944
SRCV 2006	9 985	10 116	25 056
CVS 2007	17 428	17 555	41 564
SRCV 2007	10 455	10 571	26 020
Transport 2007	20 096	20 243	49 367
CVS 2008	17 118	17 212	40 788
HSM 2008	24 087	24 087	66 033
SRCV 2008	10 384	10 465	25 594
CVS 2009	17 056	17 160	40 635
Patrimoine 2009	12 788	13 028	30 229
SRCV 2009	10 570	10 643	25 702
CVS 2010	16 465	16 571	38 073
SRCV 2010	11 011	11 095	26 617
BdF 2011	10 302	10 384	24 467
CVS 2011	16 894	16 987	38 836
SRCV 2011	11 325	11 403	27 156
CVS 2012	16 967	17 054	39 046
SRCV 2012	11 964	12 049	28 631
CVS 2013	14 548	14 613	33 171
Logement 2013	27 158	27 322	65 075
SRCV 2013	11 117	11 191	26 422
CVS 2014	16 372	16 445	37 713
Patrimoine 2014	11 628	11 767	27 910
SRCV 2014	11 370	11 440	26 898
CVS 2015	16 109	16 204	37 654
SRCV 2015	11 369	11 447	26 750

Tableau 4.1 – Liste des enquêtes empilées et effectifs

BdF : Budget de Famille ; CVS : Cadre de Vie et Sécurité ; SRCV : Statistiques sur les Ressources et Conditions de Vie ; HSM : Handicap-Santé volet Ménages

1. Comme cela est indiqué dans l'introduction, on parle ici des bases finales qui ne comprennent ni les déchets, ni les DOM, ni les enquêtes AES 2012 et CdT 2012.

Chapitre 5

Dictionnaire des variables

Ci-dessous figure le dictionnaire des variables de la base des TCM empilés. Il est pour le moment restreint aux principales variables d'intérêt car toutes les variables de collecte, pas nécessairement utiles pour les exploitations ont été gardées dans un premier temps. Il faudra encore faire un tri pour alléger les tables et en simplifier la compréhension, en harmonisant certaines des variables qui auraient évolué entre 2006 et 2015. Une fois ce travail effectué, un nouveau dictionnaire des variables, complet, sera diffusé.

Dictionnaire des variables
de la base des TCM empilés
sur la période 2006-2015

Remarque : en cas d'un besoin de renseignements sur une variable non documentée ici ou d'un besoin de compléments d'information sur des variables du dictionnaire, il faut contacter Loïc Vinet (loic.vinet@insee.fr), successeur de Céline Leroy, à compter du 3 septembre 2018.

Chapitre 6

Spécificités et problèmes rencontrés sur certaines enquêtes

Dans ce chapitre, sont recensés les différents problèmes ou particularités déjà identifiés dans les données pour certaines enquêtes.

Variables MER1E et PER1E

Dans les enquêtes CVS 2010 et 2011, un nombre anormal de valeurs manquantes est observé sur les variables de situation des parents, MER1E et PER1E (vit ici / vit ailleurs / est décédé / parent inconnu). Les variables drapeaux associées signalent qu'il ne s'agit pas de cas « Ne Sait Pas (NSP) », mais de cas où le champ de la variable n'a pas été rempli. Ce qui veut dire que les questions n'ont pas été posées dans ces cas-là alors qu'elles doivent normalement être posées à tous les individus du logement.
Pour l'enquête CVS 2010, en dehors des cas NSP, les valeurs de MER1E sont à blanc pour 26 299 individus et les valeurs de PER1E le sont pour 27 515 individus. Pour l'enquête CVS 2011, toujours en dehors des cas NSP, les valeurs de MER1E sont à blanc pour 27 003 individus et celles de PER1E le sont pour 28 287 individus.
Ce problème est lié à une anomalie dans le data-model du TCM en 2010 et 2011. En effet, dès lors que l'individu n'avait pas de mère potentielle [1] dans le logement, la question MER1E était passée et de même pour le père.
Cela pourra donc poser problème, notamment pour des études sur les orphelins, car l'information sur la situation des parents sera absente. Plus précisément, on ne saura pas s'ils vivent ailleurs, s'ils sont décédés ou s'ils sont inconnus.
Les autres enquêtes qui pourraient être concernées sont les enquêtes SRCV 2010 et 2011 et l'enquête BdF 2011. Cependant, on n'observe qu'un nombre limité de valeurs manquantes, non liées aux NSP, pour ces trois enquêtes (entre 11 et 16 valeurs manquantes plus précisément). Pour l'enquête SRCV, cela n'est pas surprenant car elle possède son propre data-model du TCM adapté à la réinterrogation [2]. En revanche, cela est plus surprenant de ne pas avoir le problème pour l'enquête BdF 2011 qui a normalement embarqué la même version du TCM que celle de l'enquête CVS 2010 (pour des questions de calendrier, elle a embarqué la version 2010 plutôt que 2011).
Il peut s'agir ici des aléas du versionning qui n'est peut-être pas toujours respecté, ou encore de l'existence d'adhérences entre le data-model du TCM et celui de l'enquête dans lequel on l'intègre, qui ferait qu'une même version du TCM, intégrée dans deux enquêtes différentes ne donne pas

1. Une mère potentielle est une personne du logement, de sexe féminin et ayant au moins 10 ans de plus que l'individu concerné par la question. Il en est de même pour un père potentiel, avec le sexe masculin.
2. Pour information, l'enquête SRCV embarquera la nouvelle version longitudinale du TCM à partir de 2020.

toujours les mêmes résultats.

Deux doublons de niveau ménage dans l'enquête SRCV 2012

Dans la table TCM annuelle 2012 de niveau ménage, il existe trois observations avec le même identifiant *IDENT_MEN_UNIQUE*. L'observation logement associée, dans la table de niveau logement, nous indique qu'il y a plusieurs budgets séparés dans ce logement et plus exactement trois. Dans la table de niveau individu, grâce à la matrice des liens établie entre les individus de ce logement, on peut déduire qu'il s'agit d'un couple avec leurs deux enfants qui font chacun budget séparé.
Ces trois ménages devraient donc avoir respectivement BS = 1, 2 et 3 mais ils ont chacun BS = 0 et donc ils se retrouvent avec le même identifiant ménage : 26004927000 srcv12. Dès lors, il est difficile de déterminer s'il y a réellement eu éclatement en trois budgets séparés au cours de la collecte.
C'est la raison pour laquelle rien n'a été modifié pour la livraison des tables finales. Mais si une étude de niveau ménage est réalisée, il sera à la charge de l'utilisateur, en fonction de ce qu'il souhaite étudier, soit de mofidier l'identifiant ménage des trois observations en remplaçant le dernier 0, respectivement par 1, 2 et 3, soit de considérer qu'il n'y a qu'un seul ménage et donc modifier les données de niveau ménage en conséquence pour garder une cohérence.

Vigilance à avoir sur la variable individuelle « nombre d'autres logements »

Entre 2006 et 2008, c'est la variable NAUTLOG de la table de niveau individu qui indique le nombre d'autres logements ordinaires de chaque individu.
À partir de 2009, cette dernière est renommée en NAUTLOG_IND.
Néanmoins, pour les enquêtes SRCV 2014 et 2015, c'est la variable NAUTLOG et non NAUTLOG_IND qui apparaît dans la table individu alors que le changement de nom a bien été pris en compte dans les enquêtes SRCV entre 2009 et 2013.
Il est donc normal d'avoir les deux variables NAUTLOG et NAUTLOG_IND dans les tables TCM_IND annuelles de 2014 et 2015.

Taille des identifiants logement, ménage et individu

Entre 2006 et 2008, la variable « Numéro de Fiche-Adresse », NUMFA, était sur quatre positions pour toutes les enquêtes. En 2009, elle est passée sur 6 positions pour l'enquête Patrimoine uniquement. Et à partir de 2010, elle est passée sur 6 positions pour toutes les enquêtes, sauf l'enquête SRCV, pour laquelle elle est restée sur 4 positions jusqu'en 2015.
Cela a une incidence sur la taille des identifiants. En effet, l'identifiant logement est construit à partir de la concaténation du numéro de région de gestion (RGES sur 2 positions), du numéro de la fiche-adresse (NUMFA), du numéro de sous-échantillon (SSECH sur 2 positions), du numéro de logement éclaté (LE sur 1 position) et du numéro de ménage éclaté (EC sur 1 position). L'identifiant ménage est quant à lui la concaténation de l'identifiant logement et du numéro de budget séparé (BS sur 1 position). Et enfin, l'identifiant individu est la concaténation de l'identifiant ménage et du numéro d'ordre individuel (NOI sur 2 positions). Ainsi, la variable NUMFA n'étant pas de la même taille pour toutes les enquêtes empilées, les tailles d'identifiants sont également différentes :

- entre 2006 et 2008, les identifiants logement, ménage et individu sont respectivement de taille 10, 11 et 13 ;
- à partir de 2009, ils sont respectivement de taille 12, 13 et 15 avec les deux derniers caractères à blanc pour les enquêtes CVS 2009 et SRCV 2009 à 2015.

Champ d'étude des blocs Activité professionnelle et Ressources culturelles

Dans le TCM, les blocs « F. Activité professionnelle » et « G. Ressources culturelles » (questions sur la nationalité, la formation et les diplômes) ne sont pas posés à tous les individus et le champ dépend d'options choisies par chaque concepteur d'enquête.
En effet, chaque responsable d'enquête peut choisir de poser ces deux blocs :

- uniquement aux personnes de référence du ménage (cas où le concepteur choisit OPTION_INDIV = 1) ;
- ou aux personnes âgées de AGEMIN ans ou plus (cas où le concepteur choisit OPTION_INDIV = 2).

AGEMIN est aussi une variable renseignée par le concepteur d'enquête. Elle vaut en général 14 ou 15 ans.
Remarque : pour le bloc Activité professionnelle, si OPTION_INDIV = 2, l'âge à partir duquel on pose les questions n'est pas AGEMIN mais max(15, AGEMIN).
Ainsi, il faut prendre cela en compte pour faire des études sur la profession et le diplôme à partir des TCM empilés car les champs ne sont pas harmonisés entre les différentes enquêtes. Les variables OPTION_INDIV et AGEMIN qui permettent de connaître ce champ figurent dans la table de niveau individu. Dans le cas où OPTION_INDIV = 1, il faudra également utiliser l'indicatrice d'appartenance au groupe de référence, IGREF, qui figure aussi dans la table (cf. chapitre 5 pour les modalités de ces variables).

Variables diplôme renseignées à tort

Dans les enquêtes CVS 2014 et 2015, les variables diplôme sont renseignées pour les enfants de 13 ans et moins alors que AGEMIN = 14 ans. Il s'agit d'un problème dans la chaîne aval et il ne faut pas tenir compte de cette tranche d'âge pour ces deux enquêtes.
Mais comme cela a été dit dans le point qui précède, pour chaque enquête, il ne faut s'appuyer que sur les variables OPTION_INDIV et AGEMIN ou IGREF pour savoir qui a répondu aux blocs Activité professionnelle et Ressources culturelles.

Remarque : si d'autres problèmes sont rencontrés lors de l'exploitation de la base, il faut contacter Loïc Vinet, successeur de Céline Leroy à compter du 3 septembre 2018.

Conclusion

Cette partie fait donc figure de documentation de la base. Elle contient la liste des enquêtes empilées, un dictionnaire des codes restreint aux principales variables d'intérêt des utilisateurs et un recensement des particularités à avoir en tête et des problèmes déjà repérés dans la base.
Elle pourra être complétée au fur et à mesure des retours et questions des utilisateurs et il faudra également finaliser le dictionnaire des codes, une fois que la sélection et l'harmonisation des variables entre les différentes enquêtes auront été réalisées.

Conclusion générale

Ainsi, il s'agit ici de la première livraison d'une base des TCM empilés pondérée et documentée. Elle concerne la période 2006-2015.

Nous avons vu que la stratégie de pondération choisie par la division RTI repose sur le partage des poids et les marges de l'EEC. Pour les poids de niveau logement des tables TCM annuelles, empilant les enquêtes d'une même année, la méthode optimale du partage des poids est appliquée, puis un calage niveau logement, à partir des marges EEC, est réalisé. Ensuite, dans les tables annuelles de niveau ménage, chaque ménage a le poids de son logement. Enfin, pour les poids de niveau individu, c'est la méthode classique du partage des poids qui est appliquée et elle est également suivie d'un calage de niveau individu, à partir des marges EEC. Concernant la pondération des tables TCM périodiques, empilant les tables TCM annuelles, les poids précédents sont divisés par le nombre d'années empilées puisqu'un estimateur périodique correspond à la moyenne des estimateurs annuels, donc ils sont divisés par 10 ici.
Il est néanmoins possible, pour chaque utilisateur, d'utiliser des alternatives, comme un tirage aléatoire des multi-résidents pour les poids de niveau individu, de manière à garantir des poids individus égaux entre tous les membres d'un même ménage [3] ou encore d'utiliser d'autres marges de calage. C'est pour cette raison que tous les poids intermédiaires ont été laissés dans les bases.

Quant à la documentation, elle pourra être complétée lorsque la sélection et l'harmonisation des variables auront été réalisées et le dictionnaire des variables, finalisé en conséquence. Il pourra également être opportun de compléter le recensement des spécificités et des problèmes rencontrés sur les différentes enquêtes, en fonction des retours qui seront faits par les différents utilisateurs de la base.

Cette nouvelle source sera sollicitée notamment pour des études s'insérant dans le cadre du projet Big Stat, mis en place en 2017, par Laurent Toulemon, directeur de recherches à l'Ined.

Désormais, il convient de se demander si cette livraison devra se répéter pour les périodes futures, et si c'est le cas, sur quelle durée il sera le plus approprié d'empiler. Il semblerait pour le moment qu'une période plus courte, de l'ordre de cinq années, soit préférabe de manière à justifier les hypothèses de stabilité temporelle et aussi de façon à ce qu'un même logement ait un risque plus faible d'être enquêté plusieurs fois sur la période.
Cette réflexion sur le devenir des TCM empilés pourra être menée en particulier avec les acteurs du projet Big Stat qui utiliseront la base ou encore avec la division Enquêtes et Études Démographiques de l'Insee.

3. Cette solution impliquerait néanmoins de supprimer du fichier un multi-résident sur deux.

Annexe A

Poids issu de la MGPP dans le cadre d'un sondage indirect

Nous nous plaçons ici dans le cadre général de la méthode de partage des poids présenté à la section 1.1, avec les mêmes notations.
Montrons que l'estimateur issu de la MGPP, $\hat{Y} = \sum_{i \in S_B} W_i.Y_i$, est un estimateur sans biais du total Y, et que le poids W_i associé s'écrit sous cette forme : $W_i = \sum_{j \in S_A} \theta_j . \frac{L_{j,i}}{L_i}$.

On sait que $Y = \sum_{i \in \Omega_B} Y_i$ et $L_i = \sum_{j \in \Omega_A} L_{j,i}$ = nombre total de liens que l'unité i peut avoir avec les unités j de Ω_A.
On peut encore écrire le total Y sous la forme

$$Y = \sum_{i \in \Omega_B} \frac{\sum_{j \in \Omega_A} L_{j,i}}{L_i} . Y_i = \sum_{j \in \Omega_A} \sum_{i \in \Omega_B} \frac{L_{j,i}}{L_i} . Y_i = \sum_{j \in \Omega_A} Z_j$$

où $Z_j = \sum_{i \in \Omega_B} \frac{L_{j,i}}{L_i} . Y_i$ ne dépend que de j.
Y peut donc être vu comme un nouveau total Z sur la population Ω_A. On peut alors l'estimer sans biais par son estimateur de Horvitz-Thompson :

$$\hat{Y} = \sum_{j \in S_A} \theta_j . Z_j \tag{A.1}$$

où θ_j est l'inverse de la probabilité d'inclusion de j dans S_A (i.e. poids de sondage de j dans l'échantillon S_A). L'équation A.1 s'écrit également de cette façon :

$$\hat{Y} = \sum_{j \in S_A} \theta_j \sum_{i \in \Omega_B} \frac{L_{j,i}}{L_i} . Y_i \tag{A.2}$$

Or si $i \notin S_B$ alors nécessairement $L_{j,i} = 0$ (en effet, j étant tiré dans S_A, si $L_{j,i} \neq 0$ alors c'est que nécessairement i est tiré dans S_B).
L'équation A.2 peut donc se réécrire :

$$\hat{Y} = \sum_{j \in S_A} \theta_j \sum_{i \in S_B} \frac{L_{j,i}}{L_i} . Y_i = \sum_{i \in S_B} \sum_{j \in S_A} \theta_j \frac{L_{j,i}}{L_i} . Y_i \tag{A.3}$$

D'où un estimateur sans biais de Y [1], issu de la MGPP, et le poids associé :

$$\boxed{\hat{Y} = \sum_{i \in S_B} W_i.Y_i \text{ avec } W_i = \sum_{j \in S_A} \theta_j.\frac{L_{j,i}}{L_i}}$$

1. Cela sous l'hypothèse que $L_i > 0$ comme cela a été dit dans la section 1.1.

Annexe B

Poids issu de la MGPP dans le cas de bases de sondage multiples

Nous nous plaçons ici dans le cas particulier des bases de sondage multiples présenté à la section 1.2. On considère deux bases de sondage distinctes et nous gardons les mêmes notations que dans cette section.

Montrons que l'estimateur issu de la MGPP, $\hat{Y} = \sum_{i\in S} W_i.Y_i$ est un estimateur sans biais du total Y, et que le poids W_i associé s'écrit sous cette forme $W_i = \frac{1}{L_i}(\sum_{j1\in S_1} \theta_{j1}.L_{j1,i} + \sum_{j2\in S_2} \theta^*{}_{j2}.L^*{}_{j2,i}$.

On sait que $Y = \sum_{i\in\Omega} Y_i$ et $L_i = \sum_{j1\in\Omega_1} L_{j1,i} + \sum_{j2\in\Omega_2} L^*{}_{j2,i}$ = nombre de liens qu'une unité d'échantillonnage i possède avec l'ensemble des bases.
On peut encore écrire le total Y sous la forme

$$\begin{aligned} Y &= \sum_{i\in\Omega} \frac{\sum_{j1\in\Omega_1} L_{j1,i} + \sum_{j2\in\Omega_2} L^*{}_{j2,i}}{L_i}.Y_i \\ &= \sum_{i\in\Omega}\sum_{j1\in\Omega_1} \frac{L_{j1,i}}{L_i}.Y_i + \sum_{i\in\Omega}\sum_{j2\in\Omega_2} \frac{L^*{}_{j2,i}}{L_i}.Y_i \\ &= \sum_{j1\in\Omega_1}\sum_{i\in\Omega} \frac{L_{j1,i}}{L_i}.Y_i + \sum_{j2\in\Omega_2}\sum_{i\in\Omega} \frac{L^*{}_{j2,i}}{L_i}.Y_i \end{aligned}$$

D'où

$$\boxed{Y = \sum_{j1\in\Omega_1} Z_{j1} + \sum_{j2\in\Omega_2} Z^*{}_{j2}}$$

avec $Z_{j1} = \sum_{i\in\Omega} \frac{L_{j1,i}}{L_i}.Y_i$ (ne dépend que de j1) et $Z^*{}_{j2} = \sum_{i\in\Omega} \frac{L^*{}_{j2,i}}{L_i}.Y_i$ (ne dépend que de j2).
Y peut donc être vu comme la somme des nouveaux totaux Z_1 et $Z^*{}_2$, respectivement sur les populations Ω_1 et Ω_2 ($Z_1 = \sum_{j1\in\Omega_1} Z_{j1}$ et $Z^*{}_2 = \sum_{j2\in\Omega_2} Z^*{}_{j2}$).
On peut alors l'estimer sans biais par son estimateur de Horvitz-Thompson :

$$\hat{Y} = \sum_{j1\in S_1} \theta_{j1}.Z_{j1} + \sum_{j2\in S_2} \theta^*{}_{j2}.Z^*{}_{j2} \tag{B.1}$$

où θ_{j1} et $\theta^*{}_{j2}$ sont les poids de sondage, respectivement, de j1 et j2 dans les échantillon S_1 et S_2.

L'équation B.1 s'écrit également de cette façon :

$$\hat{Y} = \sum_{j1 \in S_1} \theta_{j1} \sum_{i \in \Omega} \frac{L_{j1,i}}{L_i}.Y_i + \sum_{j2 \in S_2} {\theta^*}_{j2} \sum_{i \in \Omega} \frac{{L^*}_{j2,i}}{L_i}.Y_i \tag{B.2}$$

Or dans notre cadre, les unités d'observation i sont en fait les unités d'échantillonnage j1 ou j2, donc i renvoie à j1 ou j2 si, et seulement si i = j1 ou i = j2. Dès lors, si $i \notin S$ alors $\forall j1 \in S_1, L_{j1,i} = 0$ et ${L^*}_{j2,i} = 0$.
L'équation B.2 peut donc se réécrire :

$$\hat{Y} = \sum_{i \in S} \frac{1}{L_i} (\sum_{j1 \in S_1} \theta_{j1}.L_{j1,i} + \sum_{j2 \in S_2} {\theta^*}_{j2}.{L^*}_{j2,i}) Y_i = \sum_{i \in S} W_i.Y_i$$

D'où un estimateur sans biais de Y, issu de la MGPP, et le poids associé :

$$\boxed{\hat{Y} = \sum_{i \in S} W_i.Y_i \text{ avec } W_i = \frac{1}{L_i} (\sum_{j1 \in S_1} \theta_{j1}.L_{j1,i} + \sum_{j2 \in S_2} {\theta^*}_{j2}.{L^*}_{j2,i})}$$

Annexe C

Résolution du programme d'optimisation sous contrainte pour obtenir un estimateur optimal

Nous gardons les mêmes notations que dans la section 1.2 et nous nous plaçons dans le cas où les deux populations Ω_1 et Ω_2 sont identiques : $\Omega_1 = \Omega_2 = \Omega$.
Le total Y est estimé sans biais par deux estimateurs de Horvitz-Thompson concurrents :

$$\hat{Y_1} = \sum_{i \in S_1} \theta_i.Y_i \text{ et } \hat{Y_2} = \sum_{i \in S_2} \theta^*{}_i.Y_i$$

L'objectif est de trouver l'estimateur optimal de Y sur l'échantillon S.
Il s'écrit sous forme d'une combinaison linéaire de $\hat{Y_1}$ et $\hat{Y_2}$: $\hat{Y}_{opti} = \alpha\hat{Y_1} + \beta\hat{Y_2}$, tel que $\alpha + \beta = 1$ pour obtenir un estimateur également sans biais (en effet, dans ce cas on aura bien $\mathbb{E}\left[\hat{Y}_{opti}\right] = \alpha\mathbb{E}\left[\hat{Y_1}\right] + \beta\mathbb{E}\left[\hat{Y_2}\right] = (\alpha + \beta)Y = Y$).
Comme les échantillons S_1 et S_2 sont indépendants, $\mathbb{V}\left[\hat{Y}_{opti}\right] = \alpha^2\mathbb{V}\left[\hat{Y_1}\right] + \beta^2\mathbb{V}\left[\hat{Y_2}\right]$.
Cherchons le couple (α, β) qui permet de minimiser $\mathbb{V}\left[\hat{Y}_{opti}\right]$.
Pour cela, il faut résoudre le programme d'optimisation sous contrainte suivant :

$$\begin{cases} \min\limits_{(\alpha,\beta)} \mathbb{V}\left[\hat{Y}_{opti}\right] \\ \text{sous la contrainte } \alpha + \beta = 1 \end{cases}$$

Le lagrangien associé est $L(\alpha, \beta, \lambda) = \alpha^2\mathbb{V}\left[\hat{Y_1}\right] + \beta^2\mathbb{V}\left[\hat{Y_2}\right] - \lambda(\alpha + \beta - 1)$.
Les conditions du premier ordre, qui correspondent à l'annulation des dérivées premières sont données par le système (1) dont la résolution permet d'obtenir le couple (α, β) optimal.

$$(1) \begin{cases} \frac{\partial L(\alpha,\beta,\lambda)}{\partial \alpha} = 2\alpha\mathbb{V}\left[\hat{Y_1}\right] - \lambda = 0 \\ \frac{\partial L(\alpha,\beta,\lambda)}{\partial \beta} = 2\beta\mathbb{V}\left[\hat{Y_2}\right] - \lambda = 0 \\ \frac{\partial L(\alpha,\beta,\lambda)}{\partial \lambda} = 1 - \alpha - \beta = 0 \end{cases} \iff \begin{cases} \lambda = 2\alpha\mathbb{V}\left[\hat{Y_1}\right] = 2\beta\mathbb{V}\left[\hat{Y_2}\right] \\ \alpha + \beta = 1 \end{cases} \iff \begin{cases} \alpha = \frac{\beta\mathbb{V}\left[\hat{Y_2}\right]}{\mathbb{V}\left[\hat{Y_1}\right]} \\ \beta(1 + \frac{\mathbb{V}\left[\hat{Y_2}\right]}{\mathbb{V}\left[\hat{Y_1}\right]}) = 1 \end{cases}$$

$$\iff \begin{cases} \alpha = \frac{\beta\mathbb{V}\left[\hat{Y_2}\right]}{\mathbb{V}\left[\hat{Y_1}\right]} \\ \beta = \frac{\mathbb{V}\left[\hat{Y_1}\right]}{\mathbb{V}\left[\hat{Y_1}\right]+\mathbb{V}\left[\hat{Y_2}\right]} \end{cases} \iff \boxed{\begin{cases} \alpha = \frac{\mathbb{V}\left[\hat{Y_2}\right]}{\mathbb{V}\left[\hat{Y_1}\right]+\mathbb{V}\left[\hat{Y_2}\right]} \\ \beta = \frac{\mathbb{V}\left[\hat{Y_1}\right]}{\mathbb{V}\left[\hat{Y_1}\right]+\mathbb{V}\left[\hat{Y_2}\right]} \end{cases}}$$

On retrouve bien les expressions invoquées dans la section 1.2.

Bibliographie

[1] TOULEMON L., DENOYELLE T., Contribution XI^{es} Journées de Méthodologie Statistique, *La définition des ménages dans les enquêtes françaises : comment tenir compte des multi-résidences ?*, Janvier 2012.

[2] ARDILLY P., *Les techniques de sondage*, Editions Technip, Paris, 2006.

[3] LAVALLÉE P., *Le sondage indirect : ou la méthode généralisée du partage des poids*, Editions Ellipses, Paris, 2002.

[4] LE GUENNEC J., SAUTORY O., *La macro CALMAR 2 : redressement d'un échantillon par calage sur marges*, Avril 2005.