METHODES D'ESTIMATION SUR PETITS DOMAINES POUR L'INDICATEUR AROPE (AT-RISK OF POVERTY OR SOCIAL EXCLUSION) AU NIVEAU RÉGIONAL À PARTIR DE L'ENQUÊTE SRCV

Olivier SAUTORY (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

olivier.sautory@insee.fr

Mots-clés : Petits domaines, estimation par régression, calage, indicateurs de pauvreté

Résumé

Le futur règlement européen sur les statistiques sociales (règlement Integrated European Social Statistics ou IESS), en cours de discussion au Conseil européen, va introduire des contraintes en matière de précision pour 12 indicateurs, deux d'entre eux étant définis au niveau régional (dans la nomenclature « NUTS 2 », qui correspond au périmètre des anciennes régions) : l'estimation trimestrielle du nombre de chômeurs, à partir de l'enquête emploi en continu, et l'indicateur de pauvreté AROPE (at-risk of poverty or social exclusion), à partir du dispositif Statistique sur les revenus et les conditions de vie (SRCV). Concernant cet indicateur, il faudrait augmenter significativement la taille de l'échantillon actuel pour respecter les contraintes de précision. C'est pourquoi l'utilisation de méthodes d'estimation sur petits domaines est privilégiée par l'Insee.

P. Ardilly ([2]) a proposé une méthode d'estimation, dite « synthétique » dans la théorie de l'estimation sur petits domaines, qui n'est pas fondée sur une modélisation explicite : elle utilise une technique de calage (équivalente à la méthode d'estimation par régression généralisée) de l'échantillon national sur les structures régionales associées à des variables auxiliaires bien corrélées aux différentes formes de pauvreté, et présentes dans des sources auxiliaires exhaustives (recensement de la population, fichier des revenus disponibles localisés (RDL), bénéficiaires de l'allocation de solidarité aux personnes âgées (ASPA)).

La communication revient sur l'utilisation de cette méthode, qui a l'avantage de produire des jeux de poids – ici régionaux - indépendants des variables d'intérêt (ce que réclame Eurostat), en comparant les précisions obtenues par différentes méthodes d'estimation, : calage de l'échantillon national sur des marges nationales, calage de chaque échantillon régional sur des marges régionales, calage de l'échantillon national sur des marges régionales (estimation synthétique). Les données utilisées sont celles de l'échantillon de l'enquête SRCV 2010, déjà mobilisées dans l'étude de P. Ardilly. Lorsque la précision d'un total régional est mesurée par la variance, les gains de l'estimateur synthétique par rapport aux estimateurs « directs », i.e. fondés uniquement sur les observations de la région, sont – comme attendu – très élevés. Ces gains demeurent en général significatifs lorsque la précision est mesurée par l'erreur quadratique moyenne, qui prend en compte le biais (éventuel) de l'estimateur synthétique.

Abstract

The forthcoming European regulation on social statistics (Integrated European Social Statistics (IESS)), will introduce precision constraints at regional level for the AROPE poverty indicator (at-risk of poverty or social exclusion), using the Statistics on Income and Living Conditions (SILC). Regarding this indicator, the size of the current sample should be significantly increased to meet precision constraints. This is why the use of small area estimation methods is favored by INSEE. P. Ardilly proposed a so-called "synthetic" estimation method in the theory of small area estimation, which uses a calibration technique of the national sample on regional structures associated with auxiliary variables well correlated with the different forms of poverty, and present in exhaustive auxiliary sources.

The paper goes back to the use of this method, which has the advantage of producing sets of weights - here regional - independent of the variables of interest (what is demanded by Eurostat), by comparing the precisions obtained by different estimation methods: calibration of the national sample on national margins, calibration of each regional sample on regional margins, calibration of the national sample on regional margins (synthetic estimate). The gains in precision provided by the synthetic estimator, as measured by the RMSE, are generally significant.

1. Le contexte

Le futur règlement européen sur les statistiques sociales (règlement *Integrated European Social Statistics* ou *IESS*), en cours de discussion au Conseil européen, va introduire des contraintes en matière de précision pour 12 indicateurs, deux d'entre eux étant définis au niveau régional (dans la nomenclature « NUTS 2 », qui correspond au périmètre des anciennes régions françaises) : l'estimation trimestrielle du nombre de chômeurs, à partir de l'enquête emploi en continu, et l'indicateur de pauvreté AROPE (*at-risk of poverty or social exclusion*), à partir du dispositif Statistique sur les revenus et les conditions de vie (SRCV). Concernant cet indicateur, il faudrait augmenter significativement la taille de l'échantillon actuel pour respecter les contraintes de précision. C'est pourquoi l'utilisation de méthodes d'estimation sur petits domaines est privilégiée.

Pour la publication d'indicateurs, par exemple les taux de pauvreté, Eurostat demande à disposer de poids individuels lui permettant de calculer ces indicateurs - ou toute autre statistique d'intérêt - , et pas seulement des valeurs des indicateurs que lui communiqueraient les pays. Cette exigence empêche la mise en œuvre d'un grand nombre de méthodes d'estimation sur petits domaines : en effet, les estimateurs « petits domaines » peuvent en général se mettre sous la forme d'une somme pondérée des valeurs observées, mais ces pondérations dépendent du paramètre d'intérêt estimé. L'utilisation de telles méthodes conduirait donc à produire un jeu de pondérations par variable d'intérêt, ce qui n'est bien sûr pas envisageable. La méthode mise en œuvre pour répondre aux demandes d'Eurostat doit donc consister en la mise à disposition de jeu(x) de poids ne dépendant pas des indicateurs calculés.

P. Ardilly ([2]) a proposé une méthode d'estimation, appelée « synthétique » dans la théorie de l'estimation sur petits domaines, qui n'est pas fondée sur une modélisation explicite : elle utilise une technique de calage (équivalente à la méthode d'estimation par régression) de l'échantillon national de l'enquête SRCV sur les structures régionales associées à des variables auxiliaires bien corrélées aux différentes formes de pauvreté, et présentes dans des sources auxiliaires exhaustives (recensement de la population, fichier des revenus disponibles localisés (RDL), bénéficiaires de l'allocation de solidarité aux personnes âgées (ASPA)). Cette méthode revient à produire des jeux de poids régionaux.

La communication revient sur la méthode proposée par P. Ardilly, en comparant les précisions obtenues par différentes méthodes d'estimation, dont l'estimation synthétique. Les données utilisées sont celles de l'échantillon de l'enquête SRCV 2010, déjà mobilisées dans l'étude de P. Ardilly.

2. Aspects théoriques

On rappelle dans ce paragraphe les définitions et quelques propriétés de l'estimateur par régression et de l'estimateur par calage.

2.1. Notations, données

On considère une population U = {1 … N}, un échantillon s = {1 …k …n} tiré dans U selon un plan de sondage quelconque.

Y désigne une variable d'intérêt[1], dont on cherche à estimer le total sur U : $Y = \sum_{k \in U} y_k$.

On suppose que l'on dispose de J variables auxiliaires $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$, connues sur s, et dont les totaux X_j sur U sont connus. On supposera par la suite que la variable constante égale à 1 figure toujours

[1] ***Y*** peut être quantitative, ou bien l'indicatrice associée à une modalité d'une variable qualitative : dans ce cas on cherche à estimer l'effectif total de cette modalité dans U.

parmi ces variables : ce sera la variable $\boldsymbol{X_1}$, dont le total sur U est égal à la taille N de la population, supposée donc connue.

On note x_k le vecteur contenant les valeurs des variables auxiliaires pour l'individu k de l'échantillon, et X le vecteur des totaux des variables auxiliaires :

$$x_k = \begin{pmatrix} 1 \\ x_{jk} \\ x_{Jk} \end{pmatrix}, \quad X_j = \sum_{k \in U} x_{jk}, \quad X = \sum_{k \in U} x_k = \begin{pmatrix} N \\ X_j \\ X_J \end{pmatrix}$$

On suppose les individus de s munis de poids d_k conduisant à des estimateurs (approximativement) sans biais pour les totaux des variables de l'enquête :

$$\hat{Y} = \sum_{k \in s} d_k y_k \qquad \hat{X}_j = \sum_{k \in s} d_k x_{jk} \quad j = 1 \ldots J$$

Ces poids peuvent être les poids de sondage (i.e. les inverses des probabilités d'inclusion), ou les poids obtenus après correction de la non-réponse totale.

2.2. Estimation par régression

Pour plus de détails on peut se reporter par exemple à Särndal et *alii* [8].

On suppose qu'il existe une relation linéaire approchée (voire très approchée) entre $\boldsymbol{Y}$ et les variables auxiliaires $\boldsymbol{X_j}$: $\forall k \in U \qquad y_k = \sum_{j=1}^{J} b_j x_{jk} + \varepsilon_k = {}^t b\, x_k + \varepsilon_k$, avec ${}^t b = \left(b_1 \ldots b_j \ldots b_J\right)$

Si on connaissait les valeurs des variables $\boldsymbol{X_j}$ et $\boldsymbol{Y}$ pour tous les individus de la population, on pourrait estimer[2] b **dans la population** par le coefficient B obtenu par la méthode des moindres carrés ordinaires (m.c.o.) :

$$B = \left(\sum_{k \in U} x_k {}^t x_k \right)^{-1} \left(\sum_{k \in U} x_k y_k \right)$$

La présence d'un terme constant dans la régression (puisque $x_{1k} = 1$ pour tout k) assure que le total de $\boldsymbol{Y}$ est égal à la somme des valeurs prédites par la régression (ou encore que la somme des résidus est nulle) :

$$Y = \sum_{k \in U} y_k = \sum_{k \in U} {}^t B\, x_k$$

soit : $$Y = {}^t B\, X$$

B étant inconnu, on peut l'estimer[3] (approximativement sans biais) **dans l'échantillon** par :

$$\hat{B} = \left(\sum_{k \in s} d_k x_k {}^t x_k \right)^{-1} \left(\sum_{k \in s} d_k x_k y_k \right)$$

Il s'agit de l'estimateur obtenu en effectuant dans s la régression de $\boldsymbol{Y}$ sur les $\boldsymbol{X_j}$ pondérée par les poids d_k.

L'**estimateur par régression** est alors égal à : $$\boxed{\hat{Y}^r = {}^t \hat{B}\, X}$$

[2] au sens « estimation du modèle »
[3] au sens « théorie des sondages »

On montre que :

- $\hat{Y}^r$ peut s'écrire sous la forme d'une somme pondérée des observations y_k :

$$\hat{Y}^r = \sum_{k \in s} w_k y_k \quad \text{avec} \quad w_k = {}^t X \left(\sum_{k \in s} d_k x_k {}^t x_k \right)^{-1} d_k x_k$$

  où les poids w_k ne dépendent que des x_k

- pour toute variable auxiliaire $\boldsymbol{X_j}$: $\hat{X}_j^r = \sum_{k \in s} w_k x_{jk} = X_j$

  i.e. avec les poids w_k l'échantillon est **calé sur les totaux X_j**

- $\hat{Y}^r$ est approximativement **sans biais**

- la variance de $\hat{Y}^r$ est fonction des résidus de la régression de $\boldsymbol{Y}$ sur $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$: plus les résidus sont petits, plus faible est la variance. L'estimation par régression est d'autant plus efficace en termes de précision que la relation linéaire est « bonne ».

2.3. Estimation par calage

Pour plus de détails on peut se reporter par exemple à Deville & Särndal [5].

L'objectif du calage est de déterminer des poids w_k proches des poids initiaux d_k qui assurent le calage de l'échantillon sur les totaux X_j, i.e. $\sum_{k \in s} w_k x_{jk} = X_j \quad \forall j$.

On définit alors une « fonction de distance » D entre les poids finaux w_k et les poids initiaux d_k, et on cherche à minimiser la somme des « distances » entre d_k et w_k sous les contraintes de calage :

$$\min_{w_k} \sum_{k \in s} D(d_k, w_k) \text{ avec } \sum_{k \in s} w_k x_k = \sum_{k \in U} x_k = X$$

Si on choisit $D(d_k, w_k) = \frac{(w_k - d_k)^2}{d_k}$, i.e. la méthode de calage dite « linéaire », on montre que **les poids w_k sont ceux de l'estimateur par régression.**

Par conséquent, si on dispose d'un programme de calage (par exemple la macro SAS Calmar, ou le package R Icarus), les poids w_k de l'estimateur par régression peuvent être récupérés en sortie de ce programme, en utilisant la méthode linéaire (paramètre M = 1). Pour toute variable d'intérêt $\boldsymbol{Y}$, l'estimateur par régression s'obtient alors comme la somme pondérée des y_k par les w_k :

$$\sum_{k \in s} w_k y_k = \hat{Y}^r = {}^t \hat{B} X$$

3. Les données - enquête SRCV et variables auxiliaires

En réponse à une demande d'Eurostat, 6 indicateurs de pauvreté sont calculés au niveau régional. On s'intéressera dans la suite à deux d'entre eux, le premier parce que c'est un indicateur de pauvreté très couramment utilisé, le second parce que c'est l'indicateur pour lequel les contraintes de précision au niveau régional s'appliqueront.

1. le taux de risque de pauvreté (*at risk of poverty rate)*

Ce taux est défini comme la proportion d'individus ayant un niveau de vie inférieur à 60% du niveau de vie médian national (appelé seuil de risque de pauvreté). Le niveau de vie est défini de la façon suivante : tous les individus d'un même ménage ont le même niveau de vie, égal au revenu

disponible du ménage, après transferts sociaux, divisé par le nombre d'unités de consommation (u.c.) du ménage[4].

2. le taux de risque de pauvreté ou d'exclusion sociale, appelé indicateur AROPE (*at risk of poverty or social exclusion rate*)

Ce taux est défini comme la proportion d'individus étant dans au moins l'une des situations suivantes :

- avoir un revenu disponible inférieur au seuil de risque de pauvreté (i.e. être dans la situation de pauvreté mesurée par l'indicateur précédent) ;
- être dans un état de « privation matérielle aigüe », qui correspond à l'incapacité forcée à couvrir les dépenses liées à au moins quatre parmi les neuf éléments suivants : paiement du loyer, d'un emprunt hypothécaire ou des factures d'électricité ; chauffage adapté au logement ; dépenses imprévues ; consommation régulière de viande ou d'une autre source de protéines ; vacances ; téléviseur ; réfrigérateur ; voiture ; téléphone ;
- vivre dans un ménage à faible intensité de travail, i.e. un ménage dont les membres en âge de travailler ont travaillé à moins de 20 % de leur potentiel au cours des 12 mois précédents.

Dans la suite, U désignera une population **de ménages** k (nationale ou régionale), y_k le nombre d'individus « pauvres » du ménage k (au sens des indicateurs 1. ou 2. définis ci-dessus), n_k le nombre d'individus du ménage k.

$Y = \sum_{k \in U} y_k$ désigne le nombre total d'individus pauvres

$\theta = \sum_{k \in U} y_k \,/\, \sum_{k \in U} n_k$ désigne le taux de pauvreté

Comme on le verra dans le § suivant, les techniques d'estimation mises en œuvre nécessitent la connaissance d'une information auxiliaire censée « expliquer » la pauvreté. Plus précisément, pour chaque variable auxiliaire mobilisée, on doit connaître :

- sa valeur x_k pour tout ménage k de l'échantillon national s_{NAT} ;
- son total X_{NAT} sur la population nationale U_{NAT} ;
- son total X_{REG} sur toute population régionale U_{REG}.

Les variables auxiliaires proviennent de trois sources d'information.

1. <u>Le recensement de la population</u>

Données individuelles :
sexe, 6 tranches d'âge, 4 niveaux de diplôme, 5 groupes de nationalité, 11 catégories sociales.

Données ménages/logement :
appartenance à une ZUS (ou non), 3 tranches d'unité urbaine, 5 types de ménage, statut de locataire HLM (ou non).

[4] Le 1er adulte compte pour une u.c., les autres personnes de 14 ans ou plus pour 0.5 u.c., les enfants de moins de 14 ans pour 0.3 u.c.

Les variables de niveau ménage sont toutes des variables qualitatives. Les variables individuelles sont agrégées au niveau ménage : on calcule pour chaque ménage le nombre d'hommes, le nombre de femmes, le nombre d'individus de 14 ans ou moins, …

La liste complète des modalités de ces variables figure en annexe.

2. Le fichier Revenus Disponibles Localisés (RDL)

Ce fichier exhaustif contient des informations provenant des fichiers fiscaux, augmentées des montants (imputés) des prestations sociales. Il donne pour chaque ménage fiscal le revenu disponible après redistribution, et donc le niveau de vie des individus du ménage. Il permet ainsi de calculer des vingtiles de niveau de vie **au niveau de la France métropolitaine**, i.e. des quantiles de 5 % en 5 %, notés q_5, q_{10}, q_{15}, …, q_{95}, qui permettent de répartir la **population d'individus** en 20 sous-populations d'effectifs égaux.

En considérant deux vingtiles successifs, on définit 20 variables auxiliaires de la façon suivante. Prenons le cas des quantiles q_5 et q_{10}. On note $x_{5\text{-}10}(k)$ le nombre d'individus du ménage k qui ont un niveau de vie compris entre q_5 et q_{10}. À noter que tous les individus d'un ménage ayant le même niveau de vie, $x_{5\text{-}10}(k)$ vaut soit n_k (nombre d'individus du ménage), soit 0. Le total $X_{5\text{-}10}$ sur la population étudiée (nationale ou régionale) est calculé à partir du fichier RDL. Par construction, le total $X_{5\text{-}10}$ sur U_{NAT} est égal au $1/20^e$ de la population nationale. Mais le total $X_{5\text{-}10}$ sur U_{REG} n'a aucune raison d'être égal au $1/20^e$ de la population régionale.

On définit également $x_{0\text{-}5}(k)$ (nombre d'individus du ménage ayant un niveau de vie inférieur à q_5), et $x_{95\text{-}100}(k)$ (nombre d'individus du ménage ayant un niveau de vie supérieur à q_{95}), et les totaux correspondants $X_{0\text{-}5}$ et $X_{95\text{-}100}$.

3. Le nombre de bénéficiaires de l'Allocation de Solidarité aux Personnes Âgées (ASPA)

On connaît pour chaque ménage répondant le nombre de bénéficiaires de l'APSA, et pour chaque région, on connaît le nombre de bénéficiaires de l'APSA résidant en ménage ordinaire.

N.B. Le calage de l'enquête nationale SRCV, comme pour de nombreuses enquêtes-ménages de l'Insee, utilise une autre source d'information auxiliaire : l'enquête emploi en continu. Mais cette source ne permet pas d'obtenir des marges régionales fiables. C'est pourquoi il a été nécessaire pour cette étude de mobiliser une autre information, exhaustive, avec l'inconvénient, pour les données du RP, d'être plus ancienne que celle provenant de l'enquête emploi.

4. Estimation d'un total régional

4.1. Les régressions nationale et régionales

Pour mettre en œuvre la technique d'estimation par régression, on fait l'hypothèse qu'il existe une relation linéaire entre la variable $\boldsymbol{Y}$ (nombre d'individus pauvres du ménage, au sens d'un certain type de pauvreté) et les variables « explicatives » de la pauvreté $\boldsymbol{X_1}$ … $\boldsymbol{X_j}$ … $\boldsymbol{X_J}$ (parmi lesquelles figure la variable constante), présentées au §3. Cette relation, traduite par le vecteur b des coefficients de la régression, va dépendre *a priori* de la (sous-)population à laquelle on s'intéresse.

Au niveau national : $$\forall k \in U_{NAT} \quad y_k = {}^t b_{NAT}\, x_k + \varepsilon_k$$

Dans la population U_{NAT}, le vecteur b_{NAT} est estimé par les m.c.o. par B_{NAT}, lui-même estimé dans l'échantillon s_{NAT} par :

$$\hat{B}_{NAT} = \left(\sum_{k \in s_{NAT}} d_k\, x_k\, {}^t x_k \right)^{-1} \left(\sum_{k \in s_{NAT}} d_k\, x_k\, y_k \right) \quad \text{(régression pondérée sur } s_{NAT}\text{)}$$

Pour une région donnée REG : $\forall k \in U_{REG} \quad y_k = {}^t b_{REG}\, x_k + \varepsilon_k$

Dans la population U_{REG}, le vecteur b_{REG} est estimé par les m.c.o. par B_{REG}, lui-même estimé dans l'échantillon s_{REG} ($= s_{NAT} \cap U_{REG}$) par :

$$\hat{B}_{REG} = \left(\sum_{k \in s_{REG}} d_k\, x_k\, {}^t x_k \right)^{-1} \left(\sum_{k \in s_{REG}} d_k\, x_k\, y_k \right) \quad \text{(régression pondérée sur } s_{REG})$$

4.2. Estimateur d'un total régional : basique

On cherche à estimer un total régional Y_{REG}, qui peut aussi être vu comme le total national de la variable $\boldsymbol{Y}\, \mathbb{1}_{U_{REG}}$: $Y_{REG} = \sum_{k \in U_{REG}} y_k = \sum_{k \in U_{NAT}} y_k\, \mathbb{1}_{U_{REG}}(k)$.

L'estimateur « basique » consiste à utiliser les poids initiaux d_k :

$$\hat{Y}^{bas}_{REG} = \sum_{k \in s_{NAT}} d_k\, y_k\, \mathbf{1}_{s_{REG}}(k) = \sum_{k \in s_{REG}} d_k\, y_k$$

Cet estimateur est :

- (approximativement) sans biais,
- de variance dépendant de la taille n_{REG} du sous-échantillon s_{REG}.

4.3. Estimateur par régression - calage national

Comme on l'a vu au §2.3., utiliser l'estimation par régression pour estimer un total sur U_{NAT} équivaut à utiliser les poids résultant du **calage de l'échantillon national s_{NAT} sur les marges nationales X_{NAT}**, avec la méthode linéaire. On note w_k^{NAT} les poids obtenus, qui vérifient : $\sum_{k \in s_{NAT}} w_k^{NAT}\, x_{jk} = X_{j\,NAT} \quad \forall\, j = 1 \ldots J$

L'estimateur par régression est égal à : $\hat{Y}^{r}_{REG} = \sum_{k \in s_{NAT}} w_k^{NAT}\, y_k\, \mathbb{1}_{s_{REG}}(k) = \sum_{k \in s_{REG}} w_k^{NAT}\, y_k$

Il est approximativement sans biais, de variance dépendant de n_{REG} et des résidus de la **régression dans s_{NAT}** de la variable $\boldsymbol{Y}\, \mathbb{1}_{U_{REG}}$ sur les variables auxiliaires $\boldsymbol{X_1} \ldots \boldsymbol{X_j} \ldots \boldsymbol{X_J}$.

Il est en général plus précis que l'estimateur « basique », grâce au calage sur une information auxiliaire liée à la variable d'intérêt : $V\left(\hat{Y}^{r}_{REG}\right) < V\left(\hat{Y}^{bas}_{REG}\right)$

N.B. L'estimateur par régression du total de $\boldsymbol{Y}$ sur U_{NAT} est : $\hat{Y}^{r}_{NAT} = \sum_{k \in s_{NAT}} w_k^{NAT}\, y_k = {}^t \hat{B}_{NAT}\, X_{NAT}$.

On a l'égalité : $\hat{Y}^{r}_{NAT} = \sum_{REG} \hat{Y}^{r}_{REG}$.

4.4. Estimateur par régression spécifique à la région - calage régional

On se place maintenant au sein de la population U_{REG}. Utiliser l'estimation par régression pour estimer un total sur U_{REG} équivaut à utiliser les poids résultant du **calage de l'échantillon régional s_{REG} sur les marges régionales X_{REG}**, avec la méthode linéaire. On note w_k^{REG} les poids obtenus, qui vérifient : $\sum_{k \in s_{REG}} w_k^{REG}\, x_{jk} = X_{j\,REG} \quad \forall\, j = 1 \ldots J$.

L'estimateur par régression spécifique à la région REG est égal à : $\hat{Y}^{r/REG}_{REG} = \sum_{k \in s_{REG}} w_k^{REG}\, y_k = {}^t \hat{B}_{REG}\, X_{REG}$

Il est approximativement sans biais, de variance dépendant de n_{REG} et des résidus de la **régression dans s_{REG}** de la variable ***Y*** sur les variables auxiliaires $\boldsymbol{X_1} \dots \boldsymbol{X_j} \dots \boldsymbol{X_J}$.

Il est en général plus précis que l'estimateur par régression national : $V\left(\hat{Y}_{REG}^{r/REG}\right) < V\left(\hat{Y}_{REG}^{r}\right)$,

car la régression est en général « meilleure » que dans le cas précédent (et l'information auxiliaire utilisée, i.e. les marges régionales, est plus pertinente que les marges nationales).

4.5. Estimateur synthétique

Les trois méthodes présentées précédemment sont des méthodes d'estimation dite « directe », en ce sens qu'elles n'utilisent que les données observées sur les unité de la région (ainsi que de l'information auxiliaire). La méthode proposée dans ce paragraphe appartient à la famille des techniques d'estimation sur « petits domaines ». De façon générale, ces techniques consistent, pour estimer une quantité (par exemple un total) relative à un (petit) domaine, à utiliser de l'information disponible sur les unités observées en dehors du domaine, ainsi que de l'information auxiliaire disponible au niveau des domaines ou des unités. Ceci se fait généralement grâce à une modélisation, implicite ou explicite, qui relie la variable d'intérêt aux variables auxiliaires. Parmi ces techniques, les méthodes d'estimation « synthétique » reposent sur l'hypothèse d'égalité entre un (ou plusieurs) paramètre(s) défini sur le domaine et le(s) paramètre(s) de même nature défini(s) sur la population globale (ou sur un domaine beaucoup plus vaste).

L'hypothèse posée ici est que la relation entre la « variable de pauvreté » (i.e. le nombre d'individus pauvres d'un ménage) et les variables explicatives de la pauvreté ne dépend pas de la région :

Hypothèse (H) : $\mathbf{B_{NAT} = B_{REG}}$ pour toute région REG

Sous cette hypothèse, le total régional Y_{REG} peut s'écrire : $Y_{REG} = {}^t B_{REG}\, X_{REG} \overset{(H)}{=} {}^t B_{NAT}\, X_{REG}$

Par définition l' « estimateur synthétique » est égal à : $\boxed{\hat{Y}_{REG}^{syn} \overset{déf}{=} {}^t \hat{B}_{NAT}\, X_{REG}}$

On a : $\hat{Y}_{REG}^{syn} = {}^t X_{REG}\, \hat{B}_{NAT} = {}^t X_{REG} \left(\sum_{k \in s_{NAT}} d_k\, x_k\, {}^t x_k \right)^{-1} \left(\sum_{k \in s_{NAT}} d_k\, x_k\, y_k \right) = \sum_{k \in s_{NAT}} w_{k/syn}^{REG}\, y_k$

avec $w_{k/syn}^{REG} = {}^t X_{REG} \left(\sum_{k \in s_{NAT}} d_k\, x_k\, {}^t x_k \right)^{-1} d_k\, x_k$

Cet estimateur a la même forme que l'estimateur par régression national $\hat{Y}_{NAT}^{r} = {}^t \hat{B}_{NAT}\, X_{NAT}$; de même les poids $w_{k/syn}^{REG}$ sont similaires aux poids w_k^{NAT}. La différence vient du remplacement du vecteur des totaux nationaux X_{NAT} par le vecteur des totaux régionaux X_{REG} : tout se passe donc comme si l'on effectuait un calage de l'échantillon national s_{NAT} sur les marges régionales X_{REG} , avec la méthode linéaire. Les poids $w_{k/syn}^{REG}$ vérifient : $\sum_{k \in s_{NAT}} w_{k/syn}^{REG}\, x_{jk} = X_{jREG} \quad \forall\, j = 1 \dots J$

N.B. On a l'égalité : $\hat{Y}_{NAT}^{r} = \sum_{REG} \hat{Y}_{REG}^{SYN}$ car $X_{NAT} = \sum_{REG} X_{REG}$

La variance de l'estimateur synthétique vaut : $V\left(\hat{Y}_{REG}^{syn}\right) = V\left({}^t \hat{B}_{NAT}\, X_{REG}\right) = {}^t X_{REG}\, V\left(\hat{B}_{NAT}\right) X_{REG}$

Elle dépend de la variance de $\hat{B}_{NAT}$, donc de la taille n de l'échantillon national s_{NAT}. Cette variance est en général (nettement) inférieure aux variances des estimateurs précédents.

Mais cet estimateur est **biaisé** !

Son biais vaut :

$$\text{Biais}\left(\hat{Y}_{REG}^{syn}\right)=E\left(\hat{Y}_{REG}^{syn}\right)-Y_{REG}=E\left({}^{t}\hat{B}_{NAT}\,X_{REG}\right)-Y_{REG}\cong {}^{t}B_{NAT}\,X_{REG}-{}^{t}B_{REG}\,X_{REG}={}^{t}\left(B_{NAT}-B_{REG}\right)X_{REG}$$

Il dépend donc de la différence entre B_{NAT} et B_{REG}. Il est d'autant plus faible que les régressions dans U_{NAT} et U_{REG} sont « similaires », i.e. que l'hypothèses (H) est « à peu près » vérifiée.

La précision de cet estimateur doit donc être mesurée par son **erreur quadratique moyenne** (EQM) :

$$EQM\left(\hat{Y}_{REG}^{syn}\right)=E\left(\hat{Y}_{REG}^{syn}-Y_{REG}\right)^{2}=V\left(\hat{Y}_{REG}^{syn}\right)+\text{Biais}^{2}\left(\hat{Y}_{REG}^{syn}\right)$$

L'estimation de l'EQM d'un estimateur synthétique n'est jamais aisée.

Si $\hat{Y}_{REG}$ désigne un estimateur (approximativement) sans biais de Y_{REG}, on montre (voir par exemple Rao et Molina [7]) que l'erreur quadratique moyenne s'écrit :

$$EQM\left(\hat{Y}_{REG}^{syn}\right)=E\left(\hat{Y}_{REG}^{syn}-\hat{Y}_{REG}\right)^{2}-V\left(\hat{Y}_{REG}^{syn}-\hat{Y}_{REG}\right)+V\left(\hat{Y}_{REG}^{syn}\right)$$

estimée (approximativement) sans biais par :

$$\hat{EQM}\left(\hat{Y}_{REG}^{syn}\right)=\left(\hat{Y}_{REG}^{syn}-\hat{Y}_{REG}\right)^{2}-\hat{V}\left(\hat{Y}_{REG}^{syn}-\hat{Y}_{REG}\right)+\hat{V}\left(\hat{Y}_{REG}^{syn}\right)$$

où les $\hat{V}$ désignent des variances estimées en prenant en compte le plan de sondage.

Malheureusement, cet estimateur est instable, et il peut prendre des valeurs négatives.

Rao et Molina listent plusieurs méthodes proposées par différents auteurs, fondées sur des approximations de l'EQM estimée, et/ou sur des calculs d'EQM estimées « moyennes » sur des regroupements de domaines, qui « stabilisent » ces estimations. Leur application sur les données utilisées ici ne s'est pas révélée dans un premier temps très concluante. Dans la suite, on calculera donc un majorant de cette EQM estimée : $\left(\hat{Y}_{REG}^{syn}-\hat{Y}_{REG}\right)^{2}+\hat{V}\left(\hat{Y}_{REG}^{syn}\right)$, et on prendra l'estimateur par régression spécifique $\hat{Y}_{REG}^{r/REG}$ (calage régional) comme estimateur de « référence » $\hat{Y}_{REG}$.

5. Application aux données SRCV

Les calages

Lorsque l'on utilise la méthode linéaire, certains poids obtenus peuvent être négatifs. C'est ce qui se passe en effet aussi bien pour les calages de l'échantillon national sur les marges régionales[5] (estimateur synthétique) que pour les calages des échantillons régionaux (estimateurs spécifiques). L'existence de poids négatifs n'est pas gênante en soi, mais elle peut l'être pour un utilisateur peu familiarisé avec l'usage de poids négatifs (qui demande d'ailleurs de prendre certaines précautions si l'on utilise un logiciel[6]).

Par ailleurs, lorsque nous avons réalisé les calages spécifiques régionaux, il a fallu opérer quelques regroupements de modalités des variables auxiliaires, car certaines modalités n'étaient prises que par un tout petit nombre de ménages de l'échantillon régional (voire aucun), rendant le calage très difficile (voire impossible).

[5] avec un nombre de poids négatifs allant de 135 (sur un total de 11 044 répondants) pour Rhône-Alpes à plus de 1 800 pour l'Île-de-France et la Corse)

[6] Par exemple SAS élimine de tout calcul les observations munies de poids négatifs

Les calculs de précision

Pour évaluer la précision des estimateurs, nous n'avons pas pris en compte le plan de sondage - très complexe, notamment car il s'agit d'un panel - de l'enquête SRCV. L'idée est que les gains[7] relatifs de précision entre les différents estimateurs, obtenus avec les approximations du plan de sondage utilisées, ne devraient pas trop dépendre de ces approximations.

La procédure SURVEYMEANS de SAS a été utilisée pour estimer les variances des différents estimateurs, avec la prise en compte des poids (instruction WEIGHT). La formule d'estimation de variance programmée dans SAS résulte de l'assimilation sondage sans remise ≈ sondage avec remise.

À titre de comparaison, nous avons également utilisé une formule approchée proposée par J.-C. Deville, que l'on trouve par exemple dans Caron, Deville, Sautory [4] :

$$\hat{V}\left(\hat{Y}_\pi\right) = \frac{n}{n-1}\sum_{k\in s}(1-\pi_k)\left(\frac{y_k}{\pi_k} - D_2\right)^2$$

où D_2 est la moyenne des y_k / π_k pondérée par les $(1-\pi_k)$.

Les résultats numériques sont très similaires.

Les poids initiaux d_k utilisés pour l'estimateur « basique » sont les poids contenus dans le fichier de diffusion de l'enquête, après correction de la non-réponse et calage[8].

La procédure SURVEYREG de SAS a été utilisée pour estimer la matrice de variance de $\hat{B}_{NAT}$, avec l'instruction WEIGHT.

Les indicateurs suivants ont été calculés :

- Gain de précision, en termes d'écart-type estimé, de l'estimateur par régression spécifique par rapport à l'estimateur par régression :

$$\frac{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r/REG}\right)}{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r}\right)} \times 100 \quad \text{(a)}$$

- Gain de précision, en termes d'écart-type estimé, de l'estimateur synthétique par rapport à l'estimateur par régression :

$$\frac{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{SYN}\right)}{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r}\right)} \times 100 \quad \text{(b)}$$

- Gain de précision, en termes de racine de l'EQM estimée, de l'estimateur synthétique par rapport à l'estimateur par régression :

$$\frac{\sqrt{\hat{\mathrm{EQM}}\left(\hat{Y}_{REG}^{SYN}\right)}}{\hat{\mathrm{Std}}\left(\hat{Y}_{REG}^{r}\right)} \times 100 \quad \text{(c)}$$

Les taux de pauvreté sont calculés en divisant le nombre estimé d'individus pauvres par le nombre estimé d'individus de la région (à partir de l'enquête) :

$$\theta_{REG} = \frac{\hat{Y}_{REG}}{\hat{N}_{REG}}$$

En toute rigueur, les calculs de précision - et l'évaluation des gains de précision - devraient porter sur ces taux, et non sur les effectifs d'individus pauvres. Les résultats seraient toutefois similaires à ceux

[7] ou pertes

[8] comme indiqué au §3, il s'agit d'un calage différent de celui mis en œuvre ici

présentés ici, en particulier parce que les marges de calage régionales intègrent le nombre total d'individus de la région, i.e. $\hat{N}_{REG} = N_{REG}$.

Dans les tableaux suivants, on lit pour chaque région :

- le nombre *nobs* de ménages répondants
- le taux de pauvreté calculé à partir de l'estimateur par régression national $\hat{Y}^{r}_{REG}$ (1)
- le taux de pauvreté calculé à partir de l'estimateur par régression régional $\hat{Y}^{r/REG}_{REG}$ (2)
- le taux de pauvreté calculé à partir de l'estimateur synthétique $\hat{Y}^{syn}_{REG}$ (3)
- le coefficient de variation (estimé) de l'estimateur synthétique $CV\left(\hat{Y}^{syn}_{REG}\right) = \sqrt{\hat{V}\left(\hat{Y}^{syn}_{REG}\right)} / \hat{Y}^{syn}_{REG}$
- le gain de précision en termes d'écart-type (a) défini précédemment (2)/(1)
- le gain de précision en termes d'écart-type (b) défini précédemment (3)/(1)
- le gain de précision en termes de racine de l'EQM (c) défini précédemment (3)/(1)

Tableau 1 : Taux de risque de pauvreté

Région	nobs	Taux de pauvreté (%) cal-natio (1)	cal-regio (2)	synth (3)	CV-synth %	Gains de précision (%) Écarts-types (2)/(1)	(3)/(1)	R(EQM) (3)/(1)
11	1729	10,5	12,3	13,0	3,08	53	38	79
21	288	16,7	16,1	15,0	1,50	51	6	33
22	409	22,3	15,5	14,9	1,65	17	7	17
23	286	13,9	15,1	13,5	1,36	21	3	25
24	416	10,7	11,5	12,2	1,69	28	7	24
25	271	8,8	13,1	13,6	1,88	83	13	25
26	321	13,0	13,8	12,8	1,90	36	10	42
31	788	20,9	18,4	18,9	1,53	31	12	23
41	483	19,9	14,1	14,2	1,73	36	8	9
42	297	11,2	12,0	11,7	2,11	24	10	17
43	261	15,9	11,8	13,0	1,81	31	6	32
52	775	10,2	11,7	11,5	1,81	39	13	16
53	628	14,1	11,7	11,5	1,94	37	10	13
54	361	17,5	13,8	13,8	1,83	40	7	8
72	679	16,9	12,1	13,0	1,94	22	8	31
73	512	14,9	14,1	14,1	1,84	29	12	12
74	166	18,0	16,5	14,6	1,79	36	6	39
82	907	9,2	12,4	12,2	1,78	42	18	27
83	249	14,1	13,0	13,9	1,76	25	7	27
91	444	19,3	19,8	18,7	1,53	21	6	25
93	739	12,9	16,3	15,7	1,95	51	17	33
94	35	19,6	23,3	18,9	2,17	51	5	55

On constate que les taux de pauvreté calculés après calage régional (colonne (2)) sont en général assez proches de ceux calculés à partir de l'estimateur synthétique (colonne (3)), sans que les écarts soient systématiquement positifs ou négatif, ce qui est rassurant quant à l'existence d'un biais systématique possible. Les CV de l'estimateur synthétique sont très faibles, comme attendu puisque l'on utilise les 11 044 ménages répondants de l'enquête. Il en résulte des gains de précision en termes d'écarts-types estimés très élevés (colonne (3)/(1)). Les gains en termes d'EQM estimée sont quant à eux très satisfaisants pour la plupart des régions.

Tableau 2 : Taux de risque de pauvreté ou d'exclusion sociale - AROPE

Région	nobs	Taux de pauvreté (%) cal-natio (1)	cal-regio (2)	synth (3)	CV-synth %	Gains de précision (%) Écarts-types (2)/(1)	(3)/(1)	R(EQM) (3)/(1)
11	1729	16,4	18,4	18,1	3,13	74	43	47
21	288	24,5	21,9	21,4	1,73	48	9	15
22	409	29,5	20,0	21,0	1,95	30	10	27
23	286	17,7	19,8	19,6	1,71	26	5	6
24	416	16,5	16,7	17,9	1,89	36	11	39
25	271	15,5	17,4	19,7	2,06	73	14	80
26	321	15,2	16,1	18,7	2,01	50	15	84
31	788	29,3	25,4	25,5	1,61	36	14	15
41	483	28,6	21,4	20,2	1,82	44	10	35
42	297	17,4	18,7	17,0	2,08	49	11	58
43	261	18,9	16,3	18,8	2,01	43	9	61
52	775	14,0	15,3	17,1	2,02	36	20	88
53	628	19,1	16,5	17,1	2,08	42	14	28
54	361	23,2	21,5	19,7	1,95	45	9	44
72	679	23,2	17,5	18,7	1,95	36	11	38
73	512	20,5	20,0	19,5	1,87	49	15	27
74	166	25,7	20,7	20,6	1,79	39	7	7
82	907	13,0	16,8	17,5	1,86	61	24	53
83	249	19,0	18,6	19,8	1,83	43	9	32
91	444	25,9	27,1	24,5	1,56	40	8	53
93	739	17,6	22,3	21,4	1,97	59	22	49
94	35	28,8	26,5	24,9	2,35	113	6	18

On fait les mêmes constats que pour le tableau 1 : les taux de pauvreté calculés après calage régional (colonne (2)) sont en général assez proches de ceux calculés à partir de l'estimateur synthétique (colonne (3)), sans que les écarts soient systématiquement positifs ou négatif. Les CV de l'estimateur synthétique sont très faibles, et les gains de précision en termes d'écarts-types estimés très élevés (colonne (3)/(1)). Les gains en termes d'EQM estimée sont quant à eux très satisfaisants pour la plupart des régions.

6. Conclusion et perspectives

Ces travaux exploratoires montrent que la technique de l'estimation synthétique pourrait être une réponse aux futures contraintes de précision imposées par Eurostat pour les indicateurs AROPE régionaux.

Ils doivent être complétés par la mise en œuvre de techniques permettant d'apprécier le biais de l'estimateur synthétique, ainsi qu'une éventuelle amélioration de l'estimation de son erreur quadratique moyenne.

À titre indicatif, une comparaison avec des méthodes d'estimation sur petits domaines reposant sur des modélisations explicites (ex : Fay-Herriot [6], Battese-Harter-Fuller [3]) pourra être menée, à des fins de « validation » des résultats obtenus avec l'estimation synthétique.

Il faudra étudier également la possibilité d'introduire de nouvelles variables auxiliaires, comme le nombre de bénéficiaires du Revenu de Solidarité Active ou de l'Allocation aux Adultes Handicapés.

Concernant le calendrier de mise à disposition des données du recensement, les travaux menés actuellement sur la repondération des enquêtes annuelles du recensement sont une piste à étudier pour pouvoir disposer plus rapidement qu'actuellement de données fiables à un niveau régional.

En régime de production, il faudra bien entendu prendre en compte le plan de sondage dans les calculs de précision.

Bibliographie

[1] Ardilly, P. (2006), *Panorama des principales méthodes d'estimation sur les petits domaines*, Documents de travail Insee N°M0602.

[2] Ardilly, P. (2015), *Regional estimates of poverty indicators based on a calibration technique*, Statistical working papers, Eurostat, 2015.

[3] Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). *An error components model for prediction of county crop area using survey and satellite data.* Journal of the American Statistical Association, 83, 28–36.

[4] Caron, N., Deville, J.-C. et Sautory, O. (1998) : *Estimation de précision issue de données issues d'enquêtes*. *Document méthodologique sur le logiciel de calcul de précision Poulpe*, document de travail Méthodologie Statistique n°9806, Insee.

[5] Deville, J.-C. and Särndal, C.-E. (1992). *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, Vol. 87, n° 418, pp. 376-382.

[6] Fay, R. E. and Herriot, R. A. (1979). *Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data.* Journal of the American Statistical Association, 74, 269-277.

[7] Rao, J.N.K. and Molina, I. (2013). *Small area estimation*, Wiley.

[8] Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Wiley.

Annexe : variables auxiliaires de source RP

Variables de niveau individu

CS

Modalité 1 : ('10','11','12','13')
Modalité 2 : ('21','22','23')
Modalité 3 : ('31','33','34','35','37','38')
Modalité 4 : ('42','43','44','45','46','47','48')
Modalité 5 : ('52','53','54','55','56')
Modalité 6 : ('62','63','64','65')
Modalité 7 : ('67','68')
Modalité 8 : ('69')
Modalité 9 : ('71','77','78')
Modalité 10 : ('72','74','75')
Modalité 11 : ('81','83','84','85','86')

Âge

Modalité 1 : 14 ans ou moins au 31 / 12 de l'année de collecte
Modalité 2 : de 15 ans (compris) à 29 ans (compris) au 31 / 12
Modalité 3 : de 30 ans (compris) à 39 ans (compris) au 31 / 12
Modalité 4 : de 40 ans (compris) à 49 ans (compris) au 31 / 12
Modalité 5 : de 50 ans (compris) à 59 ans (compris) au 31 / 12
Modalité 6 : 60 ans ou plus au 31 / 12

Diplôme

Modalité 1 : individus de 20 ans et moins au 31/12 de l'année de collecte
Modalité 2 : diplôme <= BEPC, et 21 ans ou plus au 31/12
Modalité 3 : diplôme > BEPC et <= BAC (ou BP ou BT), et 21 ans ou plus au 31/12
Modalité 4 : diplôme > BAC (ou BP ou BT), et 21 ans ou plus au 31/12

Nationalité

Modalité 1 : personnes ayant 15 ans ou moins au 1er janvier de l'année de collecte
Modalité 2 : français (de naissance ou non), et 16 ans ou plus au 1er janvier
Modalité 3 : européens, et 16 ans ou plus au 1er janvier
Modalité 4 : Afrique, et 16 ans ou plus au 1er janvier
Modalité 5 : Asie, Amériques, Océanie, et 16 ans ou plus au 1er janvier

Variables de niveau ménage

Tranche d'unité urbaine (TUU)
Modalité 1 : UU <= 10 000 habitants
Modalité 2 : UU de 10 000 à 100 000 habitants
Modalité 3 : UU de 100 000 habitants et plus (y compris UU de Paris)

Type de ménage

Modalité 1 : personne vivant seule
Modalité 2 : monoparental (homme ou femme seul(e) avec enfant(s))
Modalité 3 : couple sans enfant (2 personnes dans le ménage)
Modalité 4 : couple avec enfant(s) - mais pas d'autre personne dans le ménage
Modalité 5 : ménage complexe

Location en HLM (vide)

Modalité 1 : locataire ou sous-locataire d'un logement loué vide HLM
Modalité 2 : autres cas