Estimation régionale de taux de pauvreté utilisant une technique de calage

Pascal Ardilly[1]

Résumé. On propose une méthode d'estimation régionale de six taux de pauvreté basée sur l'estimateur synthétique. Concrètement, l'estimation est obtenue par un calage de l'échantillon national de l'enquête européenne SILC sur des marges régionales issues de trois sources externes. Un benchmarking est effectué après le calage. On récupère, pour chaque région, un jeu de poids utilisable pour toute variable d'intérêt corrélée avec la pauvreté, ce qui est très pratique.

Mots-clés. Petits domaines, estimateur synthétique, calage, estimation régionale

Abstract.. We propose a method of regional estimation applied to six poverty rates and using a synthetic estimate. Concretely, the estimation is produced by a calibration technique applied to the European survey SILC national sample, on regional margins build from three external datasources. A benchmarking follows the calibration step. For each region, we get a set of weights able to produce a valid estimation for any variable of interest correlated with poverty, what is very easy.

Keywords. Small area estimation, synthetic estimate, calibration, regional estimation

L'enquête européenne « Statistics on Income and Living Conditions » (SILC) est une enquête annuelle à échantillon rotatif issue d'un plan complexe, tirée dans l'échantillon-maitre de l'Insee et qui produit entre autres des indicateurs de pauvreté au niveau national. La pondération nationale, prise en l'état et appliquée à chaque région, permettrait certes d'obtenir des estimations non biaisées des taux de pauvreté régionaux, mais au prix d'une variance d'échantillonnage considérable, qui va clairement au-delà de ce que l'on peut raisonnablement accepter. En effet, si l'enquête nationale s'appuie sur un effectif de 10 500 à 11 000 ménages répondants chaque année, l'Ile de France est la seule région à obtenir plus de 1000 ménages répondants - une région « moyenne » en interrogeant avec succès quelques centaines seulement. C'est pourquoi, afin de satisfaire une demande émanant d'Eurostat de calcul d'indicateurs au niveau régional, il a été décidé d'appliquer - sur l'échantillon annuel transversal uniquement - une méthode reposant sur un modèle de comportement (technique d'estimation dite « sur petits domaines »).

La demande européenne porte sur six paramètres, en tout et pour tout. Le premier indicateur (θ_1) donne la proportion de personnes physiques dont le niveau de vie (défini comme le revenu disponible du ménage auquel la personne appartient divisé par le nombre d'unités de compte de ce ménage) est inférieur à 60 % du niveau de vie médian (at-risk-of-poverty rate). Le second indicateur (θ_2) repère les personnes physiques soumises à une privation matérielle dite « modérée », c'est-à-dire qui déclarent ne pas pouvoir couvrir au moins 3 dépenses parmi une liste de 9 dépenses citées lors de l'enquête. Le troisième indicateur (θ_3) repère les personnes physiques soumises à une privation matérielle dite « aigüe », c'est-à-dire qui déclarent ne pas pouvoir couvrir au moins 4 dépenses parmi la liste des 9 dépenses en question. Le quatrième indicateur (θ_4) donne la proportion des individus vivant dans des ménages à faible niveau d'intensité de travail. Cet indicateur diffère des précédents parce qu'il est défini sur un champ spécifique de la population : en effet, il ne concerne que les individus vivant dans des ménages qui ne sont pas constitués

[1] Insee, Département des méthodes statistiques, 165 Bd Garibaldi 69003 LYON ; pascal.ardilly@insee.fr

uniquement de personnes de 60 ans et plus, ou uniquement d'étudiants âgés de 18 à 24 ans, ou uniquement de personnes de moins de 18 ans. Le cinquième indicateur (θ_5) quantifie la proportion des individus qui sont touchés par l'un au moins des trois états de pauvreté suivants : pauvreté au sens du premier indicateur θ_1, privation matérielle aigüe, faible intensité de travail. Cet indicateur doit être considéré comme l'indicateur « phare » d'Eurostat. En particulier, des allocations de fonds européens au niveau régional seront définies en tenant compte de cet indicateur, ce qui donne une importance toute particulière à son estimation. Le sixième et dernier indicateur (θ_6) quantifie la proportion des individus qui sont touchés par les trois états de pauvreté simultanément.

1 La méthode proposée pour l'estimation régionale

L'estimation régionale des 6 taux précédents peut être envisagée de plusieurs façons - car la panoplie des techniques « petits domaines » est très vaste - mais dans tous les cas il est nécessaire de s'appuyer sur un modèle, c'est-à-dire une hypothèse qui relie le comportement régional à un comportement supra régional (comportement de référence). Pour faciliter la tâche, la piste exploitée à ce jour prend comme référence le comportement national. Le postulat de base est le suivant : la relation entre une variable de pauvreté donnée Y_k et un ensemble de variables explicatives X_k ne dépend pas de la région. Techniquement, cela se traduit de la façon suivante :

$$Y_k = B^t.X_k + \varepsilon_k$$

où ε_k est un résidu - supposé petit - de somme nulle sur l'ensemble de la population U, et B un vecteur inconnu estimé de manière à permettre un ajustement optimum. Cette relation s'applique au niveau national, donc de la même façon sur chaque région. En l'état, cette ré écriture de Y_k n'est pas un modèle et s'accommode fort bien de variables Y_k égales à 0 ou à 1, c'est-à-dire de variables d'intérêt qualitatives. Si on considère un domaine $D \subset U$, on peut s'attendre à avoir $\sum_{k\in D} \varepsilon_k \approx 0$. L'égalité devient vraie en espérance si D est un échantillon aléatoire simple dans U, autrement dit si D « se comporte comme U ». Cela revient à considérer que la liaison entre Y et X ne dépend pas de la géographie ou, de manière équivalente, que le coefficient B national est le même que le coefficient B régional. Le basculement stricto sensu vers un modèle s'obtient en postulant $\sum_{k\in D} \varepsilon_k = 0$, c'est écrire $\sum_{k\in D} Y_k = B^t \cdot \sum_{k\in D} X_k$, ce qui permet d'estimer $\sum_{k\in D} Y_k$ puisque les X_k sont des variables auxiliaires connues. En la circonstance, D est une région et les taux de pauvreté sont des ratios, donc des rapports de totaux. Il s'agit donc d'estimer dans un premier temps des totaux régionaux du type

$$Y_{REG} = \sum_{k\in REG} Y_k$$

où REG désigne l'intégralité de la population régionale (en ménage ordinaire). Un estimateur de ce total, dans l'esprit du modèle retenu, est

$$\hat{Y}_{REG} = \hat{B}^t \cdot \sum_{k\in REG} X_k = \hat{B}^t \cdot X_{REG}$$

Le vecteur B a une dimension p et son estimation $\hat{B}$ mobilise l'ensemble des données nationales : c'est là tout l'intérêt du modèle puisque l'échantillon national est de grande taille si bien que

l'estimateur $\hat{B}$ a une (très) faible variance.

L'estimateur $\hat{Y}_{REG}$ est connu sous le nom d'estimateur « synthétique ». Comme il repose sur un modèle de comportement, il est par essence biaisé, mais en contrepartie sa variance reste très modeste. L'appréciation du biais est évidemment délicate, puisqu'on n'a jamais moyen de connaitre la « vraie valeur », mais on dispose d'un outil graphique de validation visuelle et d'une technique simple consistant à comparer la somme des estimations régionales à l'estimation nationale directe issue de SILC. S'ajoute un argument de bon sens, car si les variables explicatives sont suffisamment diversifiées et corrélées à l'état de pauvreté - c'est assez flagrant pour la pauvreté monétaire du fait de la présence du niveau de vie comme variable explicative, ça l'est moins pour les autres aspects de la pauvreté - on peut supposer que le rôle spécifique de la géographie n'est plus significatif. Peut-on considérer deux individus résidant, par exemple l'un en Rhône-Alpes et l'autre en Normandie, qui prendraient exactement les mêmes valeurs sur chacune des variables X_k explicatives et imaginer que la « pauvreté » de l'un sera significativement différente de la pauvreté de l'autre ? Le risque résiduel à ce niveau est constitué par une éventuelle non prise en compte dans le modèle d'une variable explicative majeure de la pauvreté, dont la corrélation avec l'ensemble des autres variables explicatives n'est pas très forte[2] et pour laquelle il faudrait, en plus, que la structure diffère significativement d'une région à l'autre[3]. En la circonstance, les variables auxiliaires retenues sont nombreuses et apparaissent potentiellement bien corrélées avec la situation de pauvreté. Néanmoins, il n'est pas difficile d'imaginer des facteurs explicatifs non pris en compte (pour le moins dans cette étude) mais néanmoins influents, comme par exemple le prix de l'immobilier, voire même des prix de biens et de services soumis à un effet local. Par ailleurs, nous avons produit des estimations régionales sur deux années consécutives (revenus 2008 puis 2009), ce qui permet de repérer d'éventuelles incohérences susceptibles de remettre le modèle en cause.

Il est clair que la méthodologie de l'estimation synthétique n'est pas originale et qu'il existe des méthodes d'estimation sensiblement plus sophistiquées. Néanmoins, d'une part complication n'est pas synonyme de gain d'efficacité, d'autre part il est apparu prudent de s'appuyer sur une technique facile à comprendre et à programmer, qui ne soit pas dépendante d'une expertise particulière sur le sujet très technique de l'estimation sur petits domaines. C'est la raison pour laquelle l'estimateur synthétique a été proposé comme base de la méthodologie d'estimation.

Cet argument de la simplicité a été poussé à l'extrême au niveau de la mise en œuvre, et c'est là que réside réellement l'intérêt de cette opération du point de vue technique. En effet, l'approche de base, la plus naturelle, consiste à programmer explicitement le vecteur de coefficients $\hat{B}$. Sur le plan purement technique, cela n'est pas vraiment difficile avec l'aide de SAS, mais il y a alors deux difficultés pratiques à surmonter. D'une part il faut faire le calcul pour chaque variable d'intérêt Y (au demeurant, la liste peut grossir si de nouveaux besoins sont exprimés !), d'autre part et surtout, les utilisateurs du fichier de données SILC n'auront pas les moyens de retrouver par eux-mêmes, de manière rapide et sans risque d'erreur, les estimations régionales des taux de pauvreté diffusées par l'Insee. En particulier, ils devront repasser l'intégralité de la procédure d'estimation en effectuant leur propre régression, en récupérant par eux-mêmes au préalable les vrais totaux des variables auxiliaires, ce qui est de fait très compromis. C'est pourquoi nous avons conçu une méthode de calcul des estimateurs synthétiques qui contourne ces difficultés majeures. Il est possible de vérifier le résultat suivant, qui est absolument essentiel pour justifier notre approche : si on considère le fichier national SILC et qu'on effectue un calage de l'intégralité de ce fichier sur les marges

[2] Il faut que cette éventuelle variable « cachée et oubliée » apporte une part d'explication qui lui est propre - si elle est combinaison linéaire des autres variables, alors elle est sans impact.

[3] Sinon la composante associée de $B^t.X_{REG}$ sera constante et les régions ne seront pas différenciées.

régionales formées par les totaux X_{REG} (considérées successivement, région par région) en utilisant la méthode dite « linéaire », on produit un jeu de poids calés (propres à la région traitée) $w_k^{calé}$ qui permet de retrouver immédiatement l'estimateur synthétique. Autrement dit, quelle que soit la variable d'intérêt Y (qu'elle soit quantitative ou qualitative), on a :

$$\sum_{k \in s} w_k^{calé} \cdot Y_k = \hat{B}^t \cdot X_{REG}$$

Le poids $w_k^{calé}$, intégré dans le fichier des micro données, peut être immédiatement utilisé avec n'importe quelle variable d'intérêt Y - mais évidemment sa pertinence statistique dépend intrinsèquement de la corrélation entre Y et le vecteur X.

Un développement théorique justifie l'utilisation du calage pour produire l'estimateur synthétique. En la circonstance, c'est la macro %CALMAR qui a été utilisée. On vérifie :

- que l'utilisation de la méthode linéaire fonctionne sans problème (option M=1 de %Calmar);
- qu'il est nécessaire d'effectuer au préalable une opération de normalisation des poids afin de les ramener à des valeurs compatibles avec l'ordre de grandeur des totaux régionaux ;
- que dans les marges utilisées pour le calage, il faut nécessairement inclure la taille totale de la population des unités statistiques en jeu, dans le cas présent il s'agit de la population de ménages. Cela revient à inclure la constante dans la liste des régresseurs. Le reste des variables auxiliaires est totalement libre.
- que l'utilisation d'une méthode de calage non linéaire ne conduit plus à l'estimateur synthétique et que lorsque la variable d'intérêt est qualitative, la justification théorique du calage avec une telle classe de méthodes est plus difficile à cerner. Ces considérations ne ruinent en rien les approches alternatives au linéaire et n'empêchent pas d'avoir l'intuition que le calage par une méthode non linéaire fournit quand même des résultats corrects pour estimer des effectifs - peut-être même ayant in fine des propriétés statistiques aussi satisfaisantes qu'avec l'approche linéaire.

Il y néanmoins un point perturbateur - mais manifestement sans conséquence - qu'il convient de souligner. La méthode produit des poids négatifs, parfois en grand nombre (jusqu'à 17 % des poids environ). En situation d'estimation classique, c'est intolérable. Mais en la circonstance, au-delà du coté désagréable de l'existence de ces poids, il faut voir qu'il ne s'agit que d'instruments pratiques pour calculer un estimateur synthétique et contourner le problème du calcul explicite du $\hat{B}$: l'important est seulement d'obtenir une estimation au final qui soit numériquement la bonne ! A noter que la proportion de poids négatifs traduit l'ampleur de la différence entre la structure régionale et la structure nationale du point de vue des variables de calage. Ainsi, considérant une région donnée, si la structure X_{REG} (obtenue par exploitation des sources externes) est proche de la structure nationale X_{NAT} obtenue par exploitation directe de l'échantillon national SILC, il n'y aura que relativement peu de poids négatifs.

2 La constitution des marges régionales

L'information auxiliaire mobilisée est censée expliquer « au mieux » la pauvreté. Elle provient de trois sources : le recensement général de la population, le fichier « Revenus disponibles localisés » constitué à l'Insee et des dénombrements départementaux d'allocataires à l'« Allocation de Solidarité aux Personnes Agées » issus de fichiers individuels gérés par la CNAF. Ces sources ont été mobilisées pour l'occasion, parce qu'elles offrent une information a priori bien corrélée à la

situation de pauvreté.

Les opérations actuelles de calage de l'enquête SILC nationale s'appuient sur une partie de ces informations, mais avec une différence appréciable : la source utilisée est l'enquête Emploi en continu. Or, cette source permet effectivement la production de marges nationales, mais pas de marges régionales. Le calage sur marges régionales - donc la production d'indicateurs de pauvreté locaux - nécessite une source « habilitée » à produire des estimations régionales, et en matière sociodémographique, actuellement seul le recensement offre cette possibilité. Cela signifie que dans la perspective d'une production annuelle régulière de taux de pauvreté régionaux, il faudra accepter les délais de mise à disposition du recensement, soit pour une production de données millésimées **n**, patienter environ jusqu'à la mi **n+3**. Une alternative peut être de construire des marges régionales en empilant toutes les enquêtes ménages d'une année donnée - y compris l'enquête Emploi et a priori sans calage après la phase de correction de la non-réponse (si le cumul de taille est lui-même jugé suffisant !). Les informations auxiliaires listées ci-dessous relèvent de concepts sociodémographiques qui apparaissent relativement simples et on peut espérer qu'il n'y a pas trop d'hétérogénéité entre les enquêtes, mais néanmoins cette opération n'a jamais été tentée et elle resterait très audacieuse.

a) Le recensement

Les variables auxiliaires sélectionnées sont les suivantes :

- le sexe
- l'âge (6 modalités)
- le diplôme (4 modalités)
- la nationalité (5 modalités)
- la catégorie sociale - CS (11 modalités)
- l'appartenance à une Zone Urbaine Sensible - ZUS (ou non)
- la tranche d'unité urbaine (3 modalités)
- le type de ménage (5 modalités)
- le statut de locataire en HLM (ou non)

b) Le fichier Revenus Disponibles Localisés (RDL)

Il s'agit d'un fichier exhaustif constitué annuellement par l'Insee et reprenant des informations provenant des fichiers fiscaux, en y ajoutant des montants de prestation. Actuellement, les prestations sont imputées par le pôle 'Revenus fiscaux' de l'Insee. Ce fichier a permis de produire, pour chaque région et France entière, des vingtiles de niveau de vie (quantiles de 5 % en 5 %). Le niveau de vie est une variable ménage, libellée en euros, qui est reportée à l'identique sur chaque individu, et que l'on définit comme le rapport du revenu disponible total du ménage au nombre d'unités de consommation comptées dans le ménage. Pour chaque région, on dispose donc de 19 vingtiles. Chaque vingtile est défini par rapport à la distribution du niveau de vie dans la population d'individus (et non de ménages). Si on considère deux vingtiles successifs, par définition la marge X_{REG} est égale au vingtième de la population régionale en nombre d'individus physiques (tous âges confondus). La pondération calée doit donc faire en sorte qu'à partir du fichier national SILC, ayant au préalable repéré les individus dont le niveau de vie se situe entre ces deux vingtiles, la somme des poids de ces individus redonne la marge régionale X_{REG} (le poids d'un individu est égal au poids de son ménage puisque le calage s'effectue au niveau ménage et que les variables d'intérêt sont des variables 'ménage' affectées uniformément à chaque individu du ménage).

c) Le nombre de bénéficiaires de l'Allocation de Solidarité aux Personnes Agées (ASPA)

Il a été possible d'obtenir, pour chaque région, le nombre de bénéficiaires de l'ASPA résidant en ménage ordinaire et en France métropolitaine. Cet effectif, au niveau national, est de 485 000 personnes en 2009 et de 489 000 personnes en 2010. Il était bien tentant de chercher à introduire en sus dans les marges les effectifs régionaux de deux prestations significativement corrélées à la pauvreté : l'allocation pour adulte handicapé (AAH) et surtout le revenu de solidarité active (RSA). Malheureusement, nous n'avons pas réussi à obtenir ces données auprès de la CNAF. Au demeurant, il paraît a priori difficile de séparer les allocations versées à des personnes résidant en ménage ordinaire des allocations versées à des personnes résidant en communauté. De plus, la CNAF publie des données régionales relatives à des allocataires au 31 décembre, ce qui est sur le plan conceptuel différent de ce que donne SILC, qui repère les allocataires « durant l'année », quelle que soit la période concernée.

On peut constater que certaines informations constituant les marges de calage sont conçues au niveau ménage, d'autres au niveau individu. Le calage a porté sur l'unité ménage k, ce qui signifie que les variables individuelles ont systématiquement été transformées en variables « ménage » (par simple addition des valeurs individuelles dans le ménage). De fait, toutes les variables X_k représentent des nombres d'individus dans le ménage k, vérifiant telle ou telle modalité. Les marges représentent donc un nombre total d'individus au niveau régional.

3 Les résultats

L'application de la méthode a donné des résultats satisfaisants sur le plan numérique. Ceux-ci ont été obtenus pour deux années consécutives : l'année SILC 2009, portant sur les revenus 2008, et l'année SILC 2010, portant sur les revenus 2009.

a) Au niveau national

Les marges utilisées pour le calage ont été produites bien entendu pour chaque région, mais également au niveau national. Il a donc été possible de caler l'échantillon national sur ces nouvelles marges. On peut constater que ces dernières incluent - à des subtilités près liées aux découpages en modalités - les marges utilisées de façon standard pour le redressement de l'enquête nationale SILC4, ce qui fait qu'on ne détruit pas le calage effectué initialement. Les estimations intéressantes portent sur les nombres de pauvres selon les différentes notions de pauvreté définies supra, puis évidemment sur les taux correspondants. En 2009, le calage national a produit 239 poids négatifs (le fichier initial comprend 10 602 ménages) et en 2010 il a produit 72 poids négatifs (sur 11 044 observations). Le tableau qui suit donne, pour l'année 2010, les estimations des effectifs de personnes 'pauvres' selon les six définitions respectives de la pauvreté, à chaque fois respectivement avant calage nouveau (colonne 'Avant') et après calage nouveau (colonne 'Après'). On rappelle que l'estimation donnée en colonne 'avant' utilise bien des poids calés, mais calés sur les marges actuellement utilisées pour la production nationale « officielle ». Ce sont ces effectifs qui sont ensuite divisés par l'estimation du nombre de personnes physiques afin de produire les taux de pauvreté.

[4] Au niveau ménage : strate d'unité urbaine, type de ménage + âge, CS et diplôme de la personne de référence ; au niveau individu : sexe et âge.

Année 2010
Effectifs de personnes en situation de pauvreté

Pauvreté au sens de θ_1		Pauvreté au sens de θ_2		Pauvreté au sens de θ_3	
Avant	Après	Avant	Après	Avant	Après
8 098 613	8 448 001	7 700 525	7 645 298	3 529 922	3 554 406

Pauvreté au sens de θ_4		Pauvreté au sens de θ_5		Pauvreté au sens de θ_6	
Avant	Après	Avant	Après	Avant	Après
4 584 882	4 393 619	11 692 708	11 920 864	924 246	881 022

Même si les ordres de grandeur ne sont jamais bousculés par le changement des variables de calage, on constate qu'il y a des mécanismes assez subtils qui peuvent créer des écarts significatifs entre la situation avant et après refonte du calage. Ici, c'est typiquement le cas du nombre de pauvres selon θ_1. Cela étant, il n'y a pas non plus nécessairement de reproductibilité des phénomènes au cours du temps.

b) Au niveau régional

Les calages sont effectués région par région : pour chaque région, on considère le fichier national d'une part, les marges régionales d'autre part, et en troisième lieu les poids nationaux préalablement normés. On cherche des nouveaux poids aussi proches que possible des poids normés, qui pondèrent l'intégralité du fichier national afin que l'on retrouve exactement les marges locales. De fait, il y a un jeu de poids par région - mais ce jeu de poids peut être utilisé pour estimer n'importe quel total au niveau régional, dès lors qu'il s'agit d'une variable corrélée à la pauvreté (plus précisément, bien expliquée par l'ensemble des variables de calage).

Aucun calage n'a échoué - même pour les régions qui ont une structure spécifique selon X - ce qui était attendu puisqu'on utilise la méthode linéaire, qui fonctionne en toutes circonstances. Comme région spécifique, il faut compter l'Île-de-France et la Corse. Pour ces deux régions, les structures nationales SILC sont vraiment éloignées ou très éloignées de la marge régionale. Malgré ces déséquilibres majeurs, pour ne pas dire spectaculaires, du fait que la fonction de calage est linéaire le calage a pu aboutir dans ces régions comme dans toutes les autres, en 2009 aussi bien qu'en 2010. En contrepartie, on a relevé un grand nombre de poids négatifs, dans toutes les régions et aussi bien en 2009 qu'en 2010 (en 2010 par exemple, on trouve 1849 poids négatifs en Ile-de-France, 755 en Basse-Normandie, mais seulement 135 en Rhône-Alpes). Il ressort clairement que l'effectif de poids négatifs reflète l'ampleur de la différence entre la structure régionale et la structure nationale. Plus une région apparaît « différente » de la France entière (du point de vue des seules variables de calage, il s'entend), plus il y a de poids négatifs.

Une méthode - probablement la plus convaincante - pour apprécier le biais du modèle consiste à comparer l'effectif national estimé à la somme des effectifs régionaux estimés. L'effectif national estimé à partir d'un échantillon de grande taille (échantillon national SILC) est jugé a priori de bonne qualité et sert donc de référence. Si les effectifs régionaux estimés par la méthode « petits domaines » n'ont pas la qualité requise, leur sommation va s'éloigner de la cible nationale. Le tableau suivant fournit l'erreur relative, en pourcentage, pour 2009 et 2010.

Biais relatif de modèle, en pourcentage

Pauvreté au sens de	2009	2010
θ_1	-0.76	4,91
θ_2	1.50	-0,58
θ_3	5.68	0,99
θ_4	-1.71	-4,23
θ_5	-0.05	2,32
θ_6	9.45	-4,16

Lecture : si on totalise les estimations régionales, le nombre de pauvres au sens de θ_1 en 2010 est égal à (1+0.0491) = 1,0491 fois le nombre total de pauvres estimé au niveau national.

Au vu des résultats, la remise en question du modèle ne se justifie pas. En particulier, le biais de l'estimateur essentiel pour Eurostat, soit $\hat{\theta}_5$, reste - au moins pour les deux années étudiées - tout à fait modeste.

Il est de bon ton que la somme des estimations régionales, pour chaque notion de pauvreté, redonne l'estimation nationale. C'est pourquoi on a appliqué une règle de trois - généralement appelée « benchmarking » - qui, à partir des estimations « petits domaines » initiales, permet d'assurer cette propriété de cohérence avec la diffusion nationale. Il a été décidé d'effectuer le benchmarking associé à chaque estimation en utilisant comme cible l'estimation nationale obtenue à partir du nouveau calage. La stratégie sur ce point n'était pas évidente. Pour la production d'indicateurs à diffuser, il y a aussi des questions de communication et de cohérence à gérer. En effet, une alternative aurait été de faire en sorte de retrouver l'effectif de personnes 'pauvres' de la statistique officielle (donc avec le calage actuel) - mais il est vrai que cette piste était beaucoup moins défendable sur le plan strictement technique et aurait enlevé de la cohérence à l'ensemble de la démarche. Après benchmarking et par construction, le biais relatif de modèle tel qu'il a été calculé supra, devient nul[5].

Le tableau ci-dessous donne les estimations régionales après benchmarking obtenues à partir de SILC 2010 (revenus 2009), c'est-à-dire les taux régionaux finaux, prêts à la diffusion. La première colonne rappelle le chiffrage 'officiel ' du taux θ_1 obtenu à partir de la seule source RDL (disponible sur le site de l'Insee). C'est le seul indicateur de pauvreté actuellement diffusé par l'Insee au niveau régional : l'estimation des cinq autres ratios représente donc une information totalement nouvelle.

On constate que le taux officiel de pauvreté RDL est (très) proche du taux $\hat{\theta}_1$, ce qui constitue une forme de validation de la pondération régionale. Cela est confirmé avec les estimations portant sur 2009. On peut penser que cette pondération réagit correctement avec les autres taux estimés $\hat{\theta}_i$. On vérifie aussi que la méthodologie d'estimation utilisée produit in fine une disparité de situations régionales assez marquée.

[5] Cet affichage serait totalement artificiel, évidemment l'appréciation du biais du modèle doit se faire avant le benchmarking.

Estimation des taux de pauvreté régionaux - année 2010

REGION	RDL 2009	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$
11	12.50	12.43	13.50	6.26	9.43	17.70	1.88
21	14.61	14.26	14.43	6.58	11.33	20.97	1.88
22	14.43	14.05	13.65	6.46	10.83	20.52	1.94
23	12.99	12.79	13.58	6.14	10.26	19.16	1.64
24	11.78	11.58	11.60	5.22	9.25	17.50	1.22
25	13.31	13.06	12.52	5.62	9.99	19.41	1.36
26	12.52	12.37	11.66	5.30	9.73	18.47	1.25
31	18.55	18.04	16.42	7.68	12.85	25.03	2.43
41	13.88	13.52	12.62	5.82	10.47	19.82	1.56
42	11.29	11.00	10.78	4.84	8.44	16.45	1.14
43	12.85	12.66	12.26	5.56	9.63	18.71	1.39
52	11.17	11.04	11.15	4.66	7.85	16.84	0.75
53	11.17	11.05	10.23	4.47	8.34	16.82	0.79
54	13.83	13.49	11.63	5.18	9.79	19.63	1.09
72	12.93	12.50	11.20	5.10	9.25	18.39	1.07
73	13.97	13.52	11.08	5.18	9.44	19.19	1.25
74	14.66	14.16	11.80	5.34	10.73	20.41	1.20
82	11.82	11.84	11.71	5.25	8.66	17.33	1.24
83	14.00	13.62	11.73	5.26	10.08	19.77	1.17
91	18.63	17.93	13.54	6.69	12.97	24.15	2.13
93	15.75	15.25	13.56	6.34	11.16	21.20	1.71
94	19.32	18.57	14.18	7.77	13.98	24.91	2.91

Le tableau suivant permet (ici sur la seule année 2010) de comparer les situations régionales selon les différents critères de pauvreté quand on attribue à chaque région un rang par critère : on attribue le rang 1 à la région la plus riche et le rang 22 à la région la plus pauvre.

Pour le futur, des pistes de progrès sont envisageables :

- on peut espérer enrichir la liste des variables de calage en mobilisant l'information localisée sur les prestations AAH et RSA, si toutefois la CNAF parvient à isoler les allocataires en collectivités ;

- on peut également espérer, même si cela paraît a priori difficile, construire de nouvelles marges autour d'une information relative au coût de la vie - en particulier ce qui touche au coût du logement ;

- il y a probablement un progrès à attendre en homogénéisant davantage les concepts de niveau de vie provenant, d'une part de SILC et d'autre part de RDL ; le projet Filosofi en cours à l'Insee devrait répondre à cette attente.

Rangs des régions selon la notion de pauvreté
Année 2010

REGION	Rang RDL	Rang $\hat{\theta}_1$	Rang $\hat{\theta}_2$	Rang $\hat{\theta}_3$	Rang $\hat{\theta}_4$	Rang $\hat{\theta}_5$	Rang $\hat{\theta}_6$
11	6	7	15	16	7	6	17
21	17	18	21	19	19	18	18
22	16	16	19	18	17	17	19
23	10	10	18	15	14	10	15
24	4	4	6	7	6	5	8
25	11	11	13	13	12	12	12
26	7	6	8	10	10	8	10
31	20	21	22	21	20	22	21
41	13	13	14	14	15	15	14
42	3	1	2	3	3	1	5
43	8	9	12	12	9	9	13
52	1	2	4	2	1	3	1
53	2	3	1	1	2	2	2
54	12	12	7	6	11	13	4
72	9	8	5	4	5	7	3
73	14	14	3	5	8	11	11
74	18	17	11	11	16	16	7
82	5	5	9	8	4	4	9
83	15	15	10	9	13	14	6
91	21	20	16	20	21	20	20
93	19	19	17	17	18	19	16
94	22	22	20	22	22	21	22

Bibliographie

[1] Rao, J.N.K. (2003), *Small Area Estimation*, Wiley.

[2] Deville, J.C., Särndal, C.E. (1992), *Calibration Estimation in Survey Sampling*, JASA, 87, 376-382.