Méthodologie d'élaboration d'une carte de pauvreté monétaire

La méthodologie développée par Elbers, Lanjouw et Lanjouw (2002, 2003) peut être exposée en trois étapes. Dans une première étape, le modèle de régression du logarithme de la dépense par tête est estimé à partir des données d'enquête, en utilisant un ensemble de variables explicatives qui sont communes à l'enquête et au recensement. Dans une deuxième étape, les paramètres de la régression sont utilisés pour prédire la dépense de tous les ménages du recensement. Enfin, une série d'indicateurs de pauvreté et d'inégalité sont construits pour différents groupes géographiques.

Bien que l'idée soit simple, sa réalisation requiert une procédure complexe permettant de prendre en compte l'autocorrélation spatiale et l'hétéroscédasticité dans le modèle de régression. Par ailleurs, le calcul des différents indicateurs utilisés et de leurs écart-types accroît considérablement la complexité de l'exercice.

Première étape

Pour commencer, on a besoin de déterminer un ensemble de variables explicatives présentent dans les deux bases de données qui remplissent certains critères de comparabilité. En effet, pour être en mesure de reproduire une carte de pauvreté compatible avec le profil de pauvreté, il apparaît important de se restreindre aux variables qui sont pleinement comparables dans le recensement et l'enquête. On commence donc par vérifier que le libellé des questions et des réponses sont bien les mêmes dans les deux questionnaires. A partir des questions sélectionnées on construit ensuite une série de variables dont on teste la comparabilité. Bien qu'il soit préférable de tester la comparabilité des distributions de chacune des variables, en pratique on retient seulement la moyenne. Afin de maximiser le pouvoir prédictif des modèles de la seconde étape, toutes les analyses sont effectuées au niveau de chacune des strates, de même que les tests de comparabilité des différentes variables à partir desquelles les modèles définitifs seront déterminés.

Deuxième étape

On estime tout d'abord le modèle de la dépense du ménage par tête en utilisant les données de l'enquête. Afin de maximiser sa précision, l'estimation du modèle est réalisée au plus bas niveau géographique pour lequel l'enquête reste représentative. Ce niveau est la strate d'échantillonnage.

Spécifions le modèle de la dépense (y_{ch}) du ménage *h* localisé en *c*, $\mathbf{x_{ch}}$ est le vecteur des variables explicatives, et u_{ch} est le terme d'erreur :

$$\ln y_{ch} = \mathrm{E}[\ln y_{ch} \mid \mathbf{x_{ch}}] + u_{ch} \qquad (1)$$

Les localités représentent des groupes de ménages définis par le plan de sondage. Elles peuvent aussi représenter des zones de dénombrement du recensement, bien que ce ne soit pas nécessairement le cas.
Si on linéarise la précédente équation, on peut modéliser le logarithme de la dépense par tête de la manière suivante :

$$\ln y_{ch} = \mathbf{x}'_{\mathbf{ch}}\boldsymbol{\beta} + u_{ch}. \qquad (2)$$

Le vecteur des perturbations $\boldsymbol{u}$ est distribué selon $F(0,\Sigma)$. Le modèle (2) est estimé par la méthode des Moindres Carrés Généralisés (MCG). Pour estimer le modèle on a d'abord besoin d'estimer la matrice de variance-covariance Σ afin de prendre en compte la possible autocorrélation spatiale (les dépenses des ménages à l'intérieur d'un même groupe sont corrélées entre elles) et l'hétéroscédasticité. Pour se faire, on spécifie le terme d'erreur comme suit :

$$u_{ch} = \eta_c + \varepsilon_{ch} \qquad (3)$$

où η_c est l'effet de localisation et ε_{ch} est la composante individuelle du terme d'erreur.

En pratique, on estime d'abord l'équation (2) par un simple MCO, puis on utilise les résidus comme des estimés pour les perturbations, notées $\hat{u}_{ch}$. On décompose alors le résidu en une composante localisation et une autre composante ménage non corrélée :

$$\hat{u}_{ch} = \hat{\eta}_c + e_{ch} \qquad (4)$$

Le terme de localisation ($\hat{\eta}_c$) est estimé par la moyenne des résidus de chaque groupe et la composante ménage (e_{ch}) est simplement déduite. L'hétéroscédasticité dans la dernière composante de l'erreur est modélisée par la régression de son carré (e_{ch}^2) sur une longue liste de variables indépendantes du modèle (2), leur carré et leurs interactions entre elles ainsi qu'avec la variable de bien-être. Un modèle logistique est utilisé pour cela.

Ces calculs de l'erreur sont utilisés pour produire deux matrices qui sont additionnées pour donner $\hat{\Sigma}$, la matrice de variance-covariance estimée du modèle (2). Cette dernière matrice permet finalement d'estimer les coefficients du modèle (2).

Troisième étape

Pour compléter la carte on associe les paramètres estimés dans la deuxième étape aux caractéristiques de chaque ménage du recensement pour prédire le log de la dépense par tête et les perturbations simulées.

Puisque la structure très complexe des perturbations a rendu le calcul de la variance des indices de bien-être trop compliquée, la technique du bootstrap est utilisée pour obtenir une mesure de la dispersion de ces indices. A partir de l'étape précédente, on peut, pour chaque ménage du recensement, simuler la valeur de l'indicateur de bien-être ($\hat{y}_{ch}^r$) à partir de l'estimation des coefficients et des termes d'erreur :

$$\hat{y}_{ch}^r = \exp(\mathbf{x}_{ch}^{'}\tilde{\beta}^r + \tilde{\eta}_c^r + \tilde{\varepsilon}_{ch}^r) \qquad (5)$$

Cette simulation est répétée 100 fois, chaque fois en retirant l'ensemble des coefficients et des termes d'erreur. La moyenne et l'écart-type des indices de bien-être simulés donnent ainsi les deux premiers moments de l'indicateur de bien-être estimé.