PRENDRE EN COMPTE L'HÉTÉROGÉNÉITÉ SPATIALE POUR CALCULER DES ESTIMATEURS - L'APPORT DE LA « GEOGRAPHICALLY WEIGHTED REGRESSION »

Jean-Michel Floch[1] & Eric Lesage[2]

[1] *Département de l'action régionale, Insee, eric.lesage@insee.fr*
[2] *Département de l'action régionale, Insee, jean-michel.floch@insee.fr*

Résumé. Dans son panorama des méthodes d'estimation sur petits domaines, P.Ardilly présente une approche par la prédiction dans laquelle l'estimateur sur le domaine est la somme des valeurs observées dans l'échantillon et la somme des valeurs prédites pour les unités statistiques qui n'appartiennent pas à l'échantillon. Dans cette approche, les poids de sondage n'interviennent pas. On s'appuie sur un modèle de comportement individuel. Les valeurs prédites sont issues d'un modèle qui relie classiquement la variable d'intérêt à des variables auxiliaires :

$$Y_i = X_i^T \beta + \epsilon_i.$$

L'estimateur sur le domaine D s'écrit alors :

$$\widetilde{Y_D} = \sum_{i \in s_D} Y_i + \sum_{i \in U_D \setminus s_D} \hat{Y}_i.$$

Si le domaine est un territoire, l'hétérogénéité spatiale peut être importante, et conduire à des variations locales dans la relation entre la variable d'intérêt et les variables auxiliaires. Le modèle explicatif repose sur les mêmes variables, mais le vecteur de coefficients β peut varier.
L'objet de la communication est donc de proposer des estimateurs des $\hat{Y}_i$ prenant en compte la situation locale, et de voir si cette méthode permet d'améliorer les résultats que l'on obtient par des méthodes classiques.
La Geographically weighted regression (GWR) est une des plus connue parmi les méthodes permettant de prendre en compte l'hétérogénéité spatiale. Elle est assez simple à exposer. On estime un modèle linéaire sur un ensemble de points, à partir des données observées dans un voisinage de ces points, la pondération des observations diminuant lorsqu'on s'éloigne du poids d'observation. Ces méthodes, proposées dans les années 1980 par des géographes se sont développées, et sont implémentées en R.
La première partie de l'exposé sera consacrée à la présentation de la GWR, de son intérêt, de ses propriétés mais aussi de ses limites.
Dans la deuxième partie, on présentera une simulation, en tirant dans le fichier des revenus fiscaux un échantillon de ménages. La variable d'intérêt sera le nombre de ménages à bas revenus. On utilisera comme variables auxiliaires le nombre de logements et le nombre de titulaires de la CMU complémentaire.
Diverses méthodes d'estimation (utilisation des poids de sondage, estimation par la prédiction sans prise en compte de l'aspect spatial, estimation utilisant la régression géographique pondérée) seront présentées.