Estimation localisée du chômage : une application des techniques d'estimation sur petits domaines

Pascal ARDILLY[1]

Lorsqu'on souhaite estimer des paramètres définis sur des populations de relativement petite taille à partir de données d'enquête obtenues par sondage, on se trouve confronté au problème de la médiocre qualité des estimations issues des méthodes classiques. C'est une conséquence mécanique de la faible taille de l'échantillon qui recoupe ces populations (appelées "domaines"), lesquelles peuvent être de « petites » aires, comme c'est le cas par exemple lorsqu'on s'intéresse aux Zones d'emploi (ZE), au nombre de 348 en France métropolitaine. Pour améliorer la précision des estimations dans ces petites aires, il est alors nécessaire d'utiliser de l'information auxiliaire obtenue à partir de sources exhaustives ou à défaut d'enquêtes par sondage de très grande taille (par exemple les enquêtes de recensement). On dispose ainsi d'un ensemble de méthodes d'estimation dites "sur petits domaines" qui sont structurées, d'une part par la nature de l'unité statistique modélisée (le domaine lui-même ou l'individu qui le compose), d'autre part par l'importance accordée au plan de sondage, et enfin par la forme proprement dite de la modélisation.

On convient généralement que l'enquête trimestrielle Emploi de l'Insee ne permet pas, en l'état, de procéder avec une précision jugée suffisante à des estimations d'effectifs de chômeurs sur des zones géographiques infra régionales. Si on souhaite obtenir une estimation trimestrielle du nombre de chômeurs BIT par ZE, les techniques d'estimation sur petits domaines offrent des perspectives intéressantes. Actuellement, l'Insee applique une méthode de ce type, basée sur une modélisation implicite qui relie le nombre de chômeurs au sens du BIT et le nombre de demandeurs d'emploi en fin de mois fourni par Pole Emploi.

Dans une perspective de recherche de gains de qualité, plusieurs modèles concurrents ont été appliqués sur les données du premier trimestre de 2007, période favorable pour disposer d'un maximum d'informations auxiliaires : calages par ZE (dans l'esprit classique du calage, c'est-à-dire sans modélisation stochastique), modèle de Fay et Herriot (qui s'appuie sur un modèle linéaire mixte), modèle linéaire optimum sans biais dit "EBLUP_B" (lequel ne tient pas du tout compte du plan de sondage, cette fois dans l'esprit de la pratique économétrique), modèle de Poisson et modèle logistique.

Une phase préalable a distingué, d'une part la constitution d'une vaste base de données auxiliaires (agrégeant 6 sources de données plus ciblées), d'autre part une pré sélection de variables explicatives. Ensuite, les différentes techniques listées ci-dessus ont été testées. Le modèle de Poisson se décline en un modèle traditionnel et un modèle mixte qui consiste à ajouter au modèle traditionnel un terme aléatoire susceptible d'expliquer les effets spécifiques locaux que les variables explicatives explicites ne parviennent pas à traduire correctement. Le modèle logistique distingue pour sa part une approche non pondérée (selon la pratique de l'économètre) et une approche pondérée (plutôt prisée par les sondeurs). S'ajoute une subdivision due au fait que l'on peut adopter une approche traditionnelle, sans effet aléatoire local, ou une approche incluant un tel effet - auquel cas on dispose d'un estimateur logistique mixte. Il est également possible, dans tous les cas, de procéder à un benchmarking final, c'est-à-dire à une ultime opération de calage des estimations par ZE sur l'estimation directe nationale issue de l'enquête Emploi.

[1] pascal.ardilly@insee.fr

L'outil informatique central mobilisé est, soit la Proc Glimmix de SAS, soit des programmes ad hoc développés au niveau européen. Pour chaque méthode appliquée, on procède à un choix de modèle (des indicateurs spécifiques sont disponibles pour nous guider) et on produit des éléments d'appréciation de la qualité, dont la nature dépend de la méthode. Il peut d'agir d'erreurs quadratiques moyennes estimées, de comparaisons effectuées avec une estimation directe agrégée (niveau national ou niveau ZEAT par exemple), ou d'éléments graphiques permettant d'apprécier la distribution des estimations "petits domaines", en particulier de juger de l'existence éventuelle d'un biais d'échantillonnage. Nous avons eu également le souci de rapprocher les estimations obtenues pour chaque méthode de l'estimation actuellement mise en œuvre par l'Insee.

L'exposé livrera les principaux enseignements de la mise en œuvre de ces différentes méthodes.