Mardi 15 mars 2005 : 11h55-12h55 ***Conférence spéciale***

Panorama des méthodes d'estimation sur petits domaines.

Pascal ARDILLY
Insee, Unité Méthodes Statistiques

Dans les enquêtes par sondage, la qualité de l'estimation d'un paramètre dépend de manière fondamentale de la taille de l'échantillon sur lequel on construit l'estimateur. Lorsqu'il s'agit d'estimer par exemple une moyenne ou un total sur une sous population (appelée « domaine ») de petite taille, l'échantillon « utile » comprend en général peu d'individus. Dans ce cas, l'estimateur classique (généralement sans biais ou peu biaisé) possède une grande ou une très grande variance. Les méthodologues, pressés par la demande de production de statistiques finement localisées aussi bien que par la demande sociale autour de populations spécifiques, ont cherché à échapper à cet inconvénient, qui pourrait paraître à première vue comme une fatalité. Ils y sont parvenus en mettant au point un ensemble de méthodes, permettant certes de diminuer la variance, mais pas sans contrepartie : l'estimation sur petits domaines repose en effet sur une modélisation des comportements, qui d'une part justifie la diminution de variance, mais d'autre part crée du biais dont l'ampleur dépend de la pertinence du modèle. L'exposé a pour objectif de passer en revue les principales méthodes d'estimation sur petits domaines, après avoir rappelé les limites que présente l'estimation « classique ».

L'idée fondamentale dans toutes ces méthodes consiste à mobiliser l'information collectée sur l'ensemble de l'échantillon pour produire des statistiques globales qui sont re utilisées, d'une façon ou d'une autre, pour estimer le paramètre local. On peut présenter différentes typologies de méthodes : l'une consiste à distinguer les méthodes à modélisation implicite des techniques de modélisation explicite.

La seconde approche se distingue de la première par le fait qu'elle utilise un ensemble de variables auxiliaires explicitement décrites dans un modèle de comportement qui relie la variable d'intérêt à ces variables auxiliaires. En général, la modélisation explicite s'appuie sur des modèles linéaires mixtes généralisés et conduit à des estimateurs qui ont certaines propriétés d'optimalité. La partie non expliquée par les variables auxiliaires se formalise au moyen d'un résidu, qui est considéré comme une variable aléatoire traduisant la caractère stochastique du modèle (donc il ne s'agit pas de l'aléa de sondage) et sur laquelle on peut être amené à formuler une hypothèse de loi (mais ce n'est pas nécessaire). Ces modèles peuvent être décrits au niveau des domaines proprement dits (un individu est un domaine) ou au niveau de l'unité de base observée lors de l'enquête (l'individu est l'unité qui compose le domaine).

La modélisation implicite est d'un abord plus facile, et apparaît peut-être plus naturelle pour le sondeur. Elle s'effectue dans un contexte un peu différent car il est possible d'exprimer le modèle en ne faisant porter les hypothèses que sur des paramètres définis respectivement sur le domaine et sur la population, lesquels paramètres ne font pas nécessairement intervenir des variables auxiliaires. Dans ce cadre, l'estimateur le plus traditionnel est l'estimateur synthétique, qui prend la forme, sous certaines conditions, d'un estimateur par le ratio ou d'un estimateur de type post stratifié. On peut aussi concevoir des estimateurs qui combinent une estimation directe traditionnelle et une estimation de type synthétique (on parle d'estimateurs composites).
Il est également possible d'introduire une composante bayésienne dans la théorie de l'estimation sur petits domaines : on aboutit alors à la classe des modèles dits « Bayésiens hiérarchiques ».