Le projet d'utilisation des données de caisse de la grande distribution alimentaire dans l'Indice des Prix à la Consommation français

Sébastien FAIVRE[1]

L'Indice des Prix à la Consommation publié chaque mois par l'INSEE repose pour une large part sur des relevés de prix réalisés sur le terrain : 160 enquêteurs effectuent ainsi chaque mois environ 160 000 relevés de prix dans 27 000 points de vente répartis dans 96 agglomérations de France métropolitaine ainsi que dans les DOM. En métropole, les prix relevés mensuellement sont rattachés à environ 1000 variétés[2] et 21 000 couples « variété/agglomération ».

Les progrès techniques permettent désormais d'enregistrer de façon exhaustive et précise l'ensemble des articles passant à la caisse des magasins de la grande distribution.

Le projet « données de caisse » vise à utiliser le potentiel important de ces données, afin d'améliorer la précision de l'indice des prix (données exhaustives sur le champ couvert), tout en modernisant les processus de collecte et en renforçant leur fiabilité.

Sur le plan méthodologique, l'utilisation des données de caisse dans le calcul de l'indice des prix pose avant tout les deux questions suivantes :

- qualité des indices de prix produits à partir des données de caisse
- faisabilité de la collecte restante sur le champ non couvert par les données de caisse

Sur le plan de la qualité des indices « données de caisse », les travaux de simulations effectués à partir d'un échantillon de données de test donnent des résultats satisfaisants.

L'objectif des simulations était de reproduire avec les données de caisses le mode de calcul actuel de l'indice des prix à partir d'un panier annuel fixe de produits.

On constate certes une forte instabilité des codes-barres[3] qui ne permettrait pas de prendre en compte l'ensemble des séries (croisement d'un code-barres et d'un point de vente) dans le calcul de l'indice, compte tenu de l'importance des remplacements à effectuer en cours d'année.

Cependant, les simulations de calcul d'indices de prix[4] montrent qu'il est possible d'obtenir un bon niveau de précision des indices avec des tailles d'échantillon réduites (de l'ordre de 1% à 2% des séries), dans le cadre de la mise en œuvre de tirages équilibrés sur des variables auxiliaires adéquates.

[1] sebastien.faivre@insee.fr

[2] Une variété regroupe un ensemble de produits « similaires », par exemple la baguette de pain, le camembert de Normandie, les jeux de société, pour lesquels il est possible de calculer un indice de prix élémentaire pour une agglomération.

[3] Taux de disparition des séries en cours d'année de 45%.

[4] 500 tirage d'échantillon effectués.

D’autre part, la comparaison entre les indices « données de caisses » et les indices issus de la collecte enquêteurs montre une proximité satisfaisante entre ces indices.

- En ce qui concerne la collecte des prix restante, le passage aux données de caisse aurait certes un impact important sur le volume d’activité des enquêteurs.

Cependant, compte tenu du faible « taux de remplissage » des agglomérations (une variété n’étant observée en moyenne que dans 21 agglomérations sur 96), il est possible de concentrer la collecte restante dans un nombre d’agglomérations plus faible (calculé afin de maintenir constant le revenu médian des enquêteurs) sans diminuer le nombre d’agglomérations d’observation pour plus de 98% des variétés.

De plus, sous l’hypothèse très raisonnable d’un doublement du nombre de séries observées sur le champ des données de caisses[5], l’échantillon d’agglomérations ainsi obtenu pour la collecte restante est supérieur à l’échantillon d’agglomérations minimal pour obtenir le niveau de précision actuel de l’indice des prix dans le cadre du modèle d’Ardilly et Guglielmetti[6] utilisé pour déterminer l’échantillon d’agglomérations optimal pour la base 1990 (échantillon en vigueur actuellement).

[5] Les hypothèses actuellement étudiées sont plutôt d’une multiplication par 10 ou par 20 du nombre de séries suivies sur le champ des données de caisses.

[6] Modèle d’échantillonnage des prix observés dans le cadre d’un sondage à deux degrés (tirage des agglomérations puis tirage des prix au sein des agglomérations) avec sondage aléatoire simple à chaque degré.