Insee
Mesurer pour comprendre

Big Data et Statistique publique

Lettre d'information
N°2 mars 2016

Les Big Données de prix, quelles opportunités ?

Le domaine où l'Insee a le plus concrètement progressé dans ses réflexions sur l'usage possible par la statistique publique de mégadonnées en provenance de la sphère privée, est celui de l'indice de prix à la consommation, avec les opportunités offertes par une exploitation des données de caisse du secteur de la grande distribution. Au-delà des dimensions méthodologiques et techniques que recouvre une telle avancée, largement abordées dans la lettre, se pose la question de l'adaptation du cadre légal, de façon à fixer les conditions dans lesquelles la statistique publique peut se voir garantir l'accès à des bases de données privées. Le projet de loi sur le numérique, actuellement en discussion au Sénat, comporte précisément un article sur ce sujet : dans sa rédaction actuelle il stipule que le Ministre chargé de l'économie, après avis du Cnis, sur la base d'une étude de faisabilité et d'opportunité rendue publique et ayant fait l'objet d'une concertation avec les personnes morales concernées, peut décider que des entreprises transmettent les informations présentes dans leurs bases de données pour répondre au besoin d'une production statistique identifiée. Le même article offre en retour des garanties aux entreprises concernées, en fixant des critères stricts en matière de finalité, de confidentialité et de sécurité de cette transmission. Nul doute que cet article de loi, s'il est voté, marquera une grande avancée pour la statistique publique. À elle de savoir s'en saisir ensuite, pour innover et se moderniser !

Fabrice Lenglart, Directeur des Statistiques Démographiques et Sociales

Des données de caisse pour le calcul de l'indice des prix à la consommation

Le versant statistique du projet Données de caisse

Le projet **«Données de caisse»** vise à prendre en compte les données quotidiennes des prix des articles vendus dans la grande distribution dans le calcul de l'indice des prix à la consommation. Les relevés de prix collectés par les enquêteurs dans les magasins seront ainsi remplacés par les données enregistrées, en interne, par les enseignes de la grande distribution lors du paiement.

Cette nouvelle source de données permet d'accroître significativement l'échantillon des produits suivis grâce aux nouvelles technologies du Big Data qui permettent de traiter un volume très important de données. Cette extension de l'échantillon se traduit par une meilleure couverture spatiale, temporelle et permet surtout de couvrir des segments de marchés (commerce équitable, produits biologiques, etc.) qui ne peuvent l'être par la collecte traditionnelle.

Les prix fournis par les données de caisse sont les prix pratiqués lors d'une vente alors que les prix relevés par les enquêteurs sont les prix affichés en magasins. Le concept des données de caisse est donc le plus juste du point de vue statistique. Cependant, comme toute nouvelle source de données, les données de caisse nécessitent d'adapter les traitements pour tenir compte des différences avec la collecte traditionnelle, sans pour autant changer les concepts sous-jacents à l'indice des prix à la consommation.

Par exemple, parce qu'un prix n'est présent dans la base que si le produit correspondant a donné lieu à une vente, utiliser les données de caisse conduit à une augmentation des cas d'absences de prix un jour donné, qu'il faut traiter statistiquement.

Autre exemple, le remplacement d'un produit suivi dans le panier lorsque celui-ci disparaît, qui est une problématique importante des indices de prix à la consommation. La connaissance des chroniques passées des caractéristiques et des prix de tous les produits du magasin disponibles dans les données de caisse permet d'améliorer la mesure statistique de la différence de qualité entre produit sortant et produit remplaçant dans le panier par comparaison des prix sur le passé.

Par pragmatisme, le périmètre du projet se limite à l'heure actuelle aux articles alimentaires industriels, aux produits d'hygiène-beauté et aux produits d'entretien de la maison vendus dans les supermarchés et hypermarchés. À terme, on peut envisager d'étendre le champ couvert par les données de caisse à d'autres biens et d'autres types d'enseignes. Il ne couvrira bien entendu jamais le commerce artisanal (boucheries, boulangeries,…).

Le projet a été lancé en 2015, à l'issue d'une phase expérimentale initiée en 2011. Il doit permettre l'utilisation de ces données dans le calcul d'indices dès 2019. Un certain nombre de travaux ont ainsi été menés par l'Insee ces dernières années en collaboration avec quatre enseignes de la grande distribution. Ils ont porté avant tout sur la qualité des données de caisse, les méthodes de calculs d'indice à partir de ces données et la question de l'effet qualité dans le cadre de remplacement de produit. Ces travaux ont donné lieu à diverses présentations, notamment lors des derniers groupes de travail d'Eurostat sur le thème, ou à Ottawa en 2015 dans le cadre de la rencontre du groupe d'experts internationaux sur les indices de prix.

Le versant informatique du projet Données de caisse

Si on caractérise usuellement les Big Data par leur volume, leur variété, et la vitesse (les trois V, voir lettre n°1), les données de caisse correspondent plutôt à la première caractéristique. Celles que l'on reçoit sont assez volumineuses sur une année : pour le périmètre du projet nous recevrons un volume de l'ordre de 40 Giga-Octets par semaine, soit au total des dizaines de milliards de lignes de vente sur une année. Les dessins des fichiers diffèrent selon l'enseigne, mais ils sont structurés, la variété des données reçues ne pose donc pas non plus de problème.

Les traitements portés par le projet sont en revanche conséquents : l'un d'eux brasse l'ensemble des données d'une année pour constituer le panier de produits, et de manière régulière les données de caisse font l'objet de nombreux appariements, puis d'agrégation en cascade pour produire les indices. Les échéances

sont fortes : les indices doivent être diffusés mensuellement. La vitesse d'exécution des traitements est donc cruciale pour le projet Données de Caisse.

Les premiers tests effectués sur les bases de données classiques ont fait apparaître un risque de lenteur des traitements et une complexité certaine pour les optimiser, peu compatibles avec les délais de production. Expertise et expérimentation à l'appui – avec l'aide en particulier du Centre d'Accès Sécurisé aux Données (CASD)–, le choix s'est porté vers les technologies Big Data, nouvelles pour l'Insee, et plus précisément sur le système Hadoop. Grâce à la répartition des données et des traitements, les performances dépendent de manière linéaire des volumes à traiter et peuvent donc être ajustées simplement.

Le projet qui est entré en phase de réalisation depuis novembre 2015 , a débuté par le développement du contrôle/chargement des données de caisse. L'équipe informatique a été accompagnée d'un prestataire pour prendre en main les nouvelles technologies.
Les technologies Big Data sont en constante évolution : nous avons porté notre choix sur des logiciels éprouvés (Pig – Hive) dont les fonctionnalités peuvent être étendues à l'aide de fonctions écrites en Java. Ces langages de haut niveau supporteront toute évolution du moteur d'exécution Hadoop susceptible de se produire.

L'utilisation des données de caisse à l'international

Seuls quelques pays européens utilisent aujourd'hui les données de caisse pour le calcul de leur indice des prix à la consommation, même si la plupart développent des projets en ce sens. Dès 2001, la Norvège a été un des premiers pays à utiliser ce type de données dans son indice. Elle a ensuite été suivie par les Pays-Bas (2002), la Suisse (2008), la Suède (2012), et tout récemment par la Belgique et le Danemark (2016). Deux autres pays devraient y recourir d'ici un an, le Luxembourg et la Pologne.

Dans ce domaine, la méthodologie du calcul de l'indice n'est pour l'heure pas harmonisée, même si l'utilisation des données de caisse dans les indices européens est autorisée. Eurostat organise depuis un certain nombre d'années des groupes de travail sur le sujet, et a souhaité faire le tour des projets des différents pays. C'est à ce titre que l'Insee a présenté ses travaux aux équipes d'Eurostat fin novembre 2015. Les discussions en cours pourraient conduire à des premières recommandations sur l'utilisation des données de caisse dans les indices de prix européens dès cette année.

Les relevés des prix par Internet

De nombreuses données de prix sont disponibles maintenant sur Internet. Pour l'indice des prix, certains prix (tarification des billets de train par exemple) sont relevés en utilisant cette information directement accessible. Avec le développement des sites de vente en ligne, on peut envisager de collecter des données de prix sur Internet automatiquement (*scraping*), en continu et à grande échelle. Le Billion Prices Project (cofondé par deux professeurs du MIT Sloan School of Management en 2006) fut l'une des premières initiatives de ce type. En 2011, BPP suivait déjà quotidiennement plus de 5 millions de produits vendus par 300 millions d'enseignes dans plus de 70 pays.

L'objectif annoncé au départ était d'utiliser les données ainsi extraites pour construire des indices de prix nationaux agrégés quotidiens. Les comparaisons menées avec l'indice officiel des prix aux États-Unis ont alimenté de nombreux débats dans la presse compte tenu de la date de parution plus tardive de l'indice officiel. Actuellement, des réflexions ont lieu dans plusieurs instituts statistiques pour utiliser ce mode de collecte pour l'indice des prix.

Ces données qui peuvent permettre à des chercheurs en économie de réaliser des analyses (par exemple sur la rigidité des prix) présentent des limites évidentes pour construire un indice de prix officiel. En particulier, les relevés par Internet ne captent pas les prix de détail, et donc les autres formes de vente, qui correspondent encore à une large part de la consommation, et dont les évolutions peuvent être différentes des prix Internet. Par ailleurs, par rapport aux données de caisse, les relevés Internet ne fournissent aucune information quant aux quantités effectivement vendues. En outre, les indicateurs du BPP ne rendent pas compte des prix d'un panier de biens qui serait représentatif de la consommation moyenne, certains de ces biens n'étant pas accessibles par Internet, comme le secteur des services par exemple.

Ce mode de collecte des données soulève par ailleurs la question juridique du respect de la propriété intellectuelle des sites «scrapés».

Pour approfondir

Les données de caisse : vers des indices de prix à la consommation à utilité constante, Patrick Sillard, Document de travail Insee n°F1305

Scanner data and quality adjustment, Isabelle Léonard, Patrick Sillard, Gaëtan Varlet and Jean-Paul Zoyem, 2015, présentation au groupe d'Ottawa, groupe international d'experts sur les indices des prix.

Actualités / Brèves

Le projet européen d'ESSnet Big Data (voir lettre n°1) est en cours de signature par Eurostat. Les travaux devraient commencer prochainement.

L'Ensae et l'Ensai ont tenu un stand au salon Big Data qui s'est déroulé les 7 et 8 mars. Le directeur du CASD a présenté les travaux menés pour mettre à disposition des chercheurs cette technologie et les essais réalisés pour le compte de l'Insee sur les données de caisse dans le guide 2015-2016 du salon. http://www.bigdataparis.com/

Agenda

Conférence du CASD: «Vos données au cœur de la data-science» le 6 avril.

Outils pour la Data Science, «Rendez-vous Méthodes et Logiciels» de la Société Française de Statistique, le 7 avril.

Ont participé à ce numéro Pierre Vernedal, Isabelle Léonard, Marie Leclair, Pascal Chevalier, Fabrice Lenglart, Stéphanie Combes, Françoise Dupont, Pauline Givord.

Cette lettre est une occasion d'informer largement et d'échanger. N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à Stéphanie Combes : stephanie.combes@insee.fr