Insee
Direction des statistiques économiques et sociales
Unité des prix à la consommation et des enquêtes ménages
Division des prix à la consommation

18, Bd Adolphe Pinard
PARIS 14

Étude de faisabilité et d'opportunité relative aux « données de caisse »

Une nouvelle source de données pour l'indice des prix à la consommation

Septembre 2016

Table des matières

L'Insee mesure chaque mois l'inflation en France : l'indice des prix à la consommation (IPC) est une mesure phare pour l'analyse économique mais aussi dans le domaine social car de nombreux contrats sont indexés sur cet indice.

Actuellement, pour le calculer, l'Insee a recours majoritairement à des enquêteurs qui observent, chaque mois, les prix des produits de consommation dans les différents commerces. Les données de caisse, données collectées par les enseignes dans leur magasin chaque fois qu'un produit est vendu et scanné à la caisse d'un magasin, sont une alternative à cette collecte traditionnelle. Dans les autres pays européens comme en France, les instituts statistiques s'intéressent à cette source de données et ont démontré qu'elle pouvait être utilisée pour calculer l'indice des prix à la consommation ; certains pays l'utilisent d'ailleurs déjà.

Dans le cadre d'une expérimentation avec des enseignes volontaires, l'Insee a pu appréhender les difficultés à manier les données de caisse (taille des bases de données, disparition et apparition de nouveaux produits, promotion, saisonnalité des produits…), les solutions à apporter mais également les gains et innovations que permettent les données de caisse. Par exemple, le volume des données améliore grandement la précision de l'indice, notamment pour des segments particuliers de la consommation (par exemple, produits éco-labellisés, bio, issus du commerce équitable) ; l'information sur l'exhaustivité des ventes et sur les chiffres d'affaires associés à la vente de chaque produit est précieuse pour s'assurer de l'absence de biais statistique (en l'absence de cette information, l'échantillonnage de l'IPC était partiellement réalisé par quotas)…

En sus de ces importantes améliorations statistiques et méthodologiques, l'accès aux données de caisse pourrait permettre à terme la production de nouvelles informations statistiques : comparaison des niveaux de prix, prix moyens, indice des prix pour de nouveaux segments de consommation…

Enfin, la substitution d'une partie de la collecte traditionnelle par les données de caisse permet de produire l'indice des prix à la consommation à un coût moindre. Du point de vue des enseignes, l'expérimentation a montré que ce changement était transparent en termes de charge : les données de caisse sont transmises à l'Insee par un tiers, une société d'études de marché à laquelle les enseignes fournissent déjà les données à des fins d'études.

Un article du projet de loi pour une république numérique, en cours d'approbation, fournit le cadre permettant la transmission au service statistique public de bases de données détenues par des personnes morales de droit privé à des fins d'enquête statistique. L'obligation de transmission des données[1] est conditionnée à la décision du ministre chargé de l'économie après une étude publique sur la faisabilité et l'opportunité d'une telle transmission et l'avis du conseil national de l'information statistique (Cnis). En contrepartie, l'Insee s'engage à assurer la confidentialité des données transmises, que ce soit dans leur transfert ou dans le cadre de leur utilisation, et à limiter strictement cette utilisation à des fins statistiques, décrites dans ce dossier.

Le présent dossier constitue l'étude de faisabilité et d'opportunité destinée au Cnis portant sur la transmission et l'utilisation des données de caisse pour le calcul de l'indice des prix à la consommation.

---

[1] *En cas de refus de la personne sollicitée, celle-ci peut encourir une amende administrative, portée à 50 000 euros en cas de récidive.*

1 - Objectifs et enjeux du projet données de caisse

1.1 - L'indice des prix à la consommation, une production phare de l'Institut qui intègre des innovations en permanence

L'indice des prix à la consommation (IPC) est l'instrument de mesure de l'inflation. Il permet d'estimer, entre deux périodes données, la variation du niveau général des prix des biens et des services consommés par les ménages sur le territoire français. C'est une mesure synthétique des évolutions de prix à qualité constante qui couvre tous les biens et services consommés sur l'ensemble du territoire national. Très utilisé à des fins économiques (déflateur d'agrégats économiques), socio-économiques (indexations), monétaires et financières (comparaisons internationales), l'IPC est publié dès la fin du mois avec une première estimation. Dans sa version harmonisée entre pays européens (IPCH), l'indice est un indicateur majeur pour la conduite de la politique monétaire dans la zone euro. Dans le cas de la France, l'IPC et l'IPCH ont des évolutions assez proches, reflet de leur proximité méthodologique. L'indice des prix à la consommation est calculé chaque mois à partir d'un échantillon de prix relevés dans un échantillon de points de vente par les enquêteurs de l'Insee. Sa production constitue l'une des obligations régaliennes de l'Insee.

L'IPC français est internationalement reconnu pour sa robustesse et sa qualité. Néanmoins, l'essor des références des articles vendus aux consommateurs du fait de la diversification des gammes (marques de distributeurs, produits à bas coût, diététiques, issus de l'agriculture biologique, du commerce équitable, etc.) et le fort développement des ventes promotionnelles à l'initiative des commerçants comme des fabricants rendent souhaitables une augmentation forte de la taille de l'échantillon des produits suivis et une connaissance simultanée des prix et des quantités vendues. Ce besoin avait été exprimé dès 2008 dans le rapport de la commission Quinet « mesurer le pouvoir d'achat des ménages », et l'est régulièrement par Eurostat, la direction statistique de la commission européenne.

L'Insee explore aujourd'hui toutes les pistes envisageables pour augmenter le volume des prix relevés (web scrapping, permettant de récupérer automatiquement des informations de prix disponibles dans certains sites web, utilisation des sources administratives, etc.). Les données de caisse font partie de ces pistes.

1.2 - Les données de caisse permettront de produire des statistiques nouvelles et d'améliorer la qualité de l'indice des prix à la consommation

L'indice des prix à la consommation est produit chaque mois à partir de la collecte de 200 000 prix dans les 30 000 points de vente de 106 unités urbaines de métropole et des DOM par plus de 200 enquêteurs de l'Insee. En complément de cette collecte terrain, 190 000 prix sont collectés centralement en bureaux par réception de fichiers, collecte internet ou web-scrapping (prix des actes médicaux, des médicaments, du transport ferroviaire ou aérien, etc.). Les données de caisse constituent une nouvelle source de données comparable aux fichiers administratifs déjà transmis.

Les avantages du recours aux données de caisse pour l'indice des prix à la consommation sont triples : (i) il est moins onéreux que la collecte traditionnelle du fait de l'économie de la collecte terrain qui fait plus que compenser des coûts d'exploitation plus importants ; (ii) le volume d'informations contenu dans les données de caisse permet la production à terme de nouvelles statistiques ou à un rythme plus fréquent qu'actuellement ; (iii) il permet une amélioration sensible

de la qualité de l'IPC.

Plus précisément, les données de caisse, par le volume d'information qu'elles contiennent, permettent de produire des indices de prix à la consommation plus précis et plus détaillés en termes de segments de marchés et de zones géographiques. Les données de caisse permettent d'envisager de produire des indices de prix régionaux, alors qu'actuellement l'IPC n'est produit que pour la France entière, la France métropolitaine et les départements d'outre-mer, faute d'un nombre d'observations suffisants. Des indices et prix moyens sur des micro-marchés qui se développent rapidement (produits éco-labellisés, bio, issus du commerce équitable…) pourront également être produits grâce à cette nouvelle source d'information. Les données de caisse permettent aussi la réalisation plus fréquente des comparaisons spatiales de prix ; ces comparaisons de niveaux de prix entre régions sont produites actuellement à un rythme quinquennal et seulement pour quelques régions au moyen d'une enquête spécifique.

En dehors des possibilités qu'offrent les données de caisse pour produire de nouvelles statistiques, les données de caisse sont une voie d'amélioration sur la qualité de l'IPC :

- ces données apportent à l'Insee la connaissance des prix de vente alors que les enquêteurs relèvent les prix affichés ; les deux peuvent différer en cas d'erreur d'affichage par exemple. Or, ce sont bien les prix effectifs qui sont à prendre en compte pour étudier l'inflation ;
- le suivi des prix dans les agglomérations de petite taille, pour lesquelles l'Insee manque d'enquêteurs, pourrait alors être pérennisé ; or, afin d'assurer la représentativité de l'indice des prix à la consommation, couvrir la totalité du territoire est une exigence européenne ;
- l'accès à l'exhaustivité des ventes permet la maîtrise du tirage aléatoire de l'échantillon des produits qui composent le panier suivi (actuellement par quotas) et du calcul de la précision de l'indice. L'absence de biais statistique peut être obtenue en fondant le calcul de l'indice sur l'exhaustivité des ventes ;
- la connaissance des chiffres d'affaires de chaque article vendu dans chaque point de vente permet de repérer rapidement les produits nouveaux et de juger au mieux de la pertinence de leur incorporation dans l'indice ; elle permet également d'associer à chaque produit son poids effectif dans la consommation des ménages. Jusqu'à présent, cette information n'était pas connue à un niveau aussi fin. Seul le poids des grands postes de la consommation était connu (les céréales pour petit déjeuner par exemple), mais pas celui d'un produit (une boîte de céréales d'une marque, taille et type donnés) ;
- la connaissance des prix des articles avant leur introduction dans le panier des biens et services suivis – en remplacement d'articles qui ne sont plus commercialisés – permet la mise en œuvre de techniques statistiques innovantes en matière de traitement de l'évolution de la qualité des articles. C'est un enjeu majeur pour les indices des prix car ceux-ci doivent mesurer l'évolution des prix à qualité constante. La disparition de produits en cours d'année et leur remplacement dans le panier de l'IPC par de nouveaux, potentiellement de qualité différente, posent alors problème au statisticien dans le cadre de la collecte traditionnelle ; observant le prix des deux produits, disparu et remplaçant, à des dates différentes, il ne sait ce qui est dû dans l'écart de prix mesuré à de l'inflation ou à une différence de qualité des produits. Les données de caisse permettent d'observer a posteriori le prix du produit disparu et de son remplaçant sur une période où les deux étaient commercialisés de manière concomitante, et d'identifier ainsi l'écart de prix propre à l'effet qualité.

1.3 – Le recours aux données de caisse dans le calcul des Indices de prix européens est une préoccupation forte

La production de nouvelles statistiques (indices de prix régionaux ou sur des micro-marchés)

comme l'augmentation de la fréquence des comparaisons spatiales de prix sont des points qui s'inscrivent dans un contexte de recherche sur le sujet au niveau européen et mondial. La France, à l'instar de nombreux autres pays, a déjà présenté un certain nombre de travaux d'étude sur ces sujets dans des réunions d'experts internationaux (groupe d'experts des prix à la consommation de l'ONU, groupe académique d'Ottawa, etc.).

Des réflexions méthodologiques sur le sujet sont également organisées par Eurostat, la direction statistique de la commission européenne. La plupart des pays européens ont engagé des projets visant à calculer une partie de leur indice des prix à partir des données de caisse. Les « workshops » organisés par Eurostat sont l'occasion d'échanger sur les aspects pratiques et méthodologiques de l'utilisation des données de caisse.

Les Pays-Bas (depuis 2002), la Norvège (2005), la Suisse (2008), la Suède (2012), la Belgique (2015) et le Danemark (2016) ont d'ores et déjà introduit les données de caisse dans le calcul de l'IPCH. Le Luxembourg et la Pologne s'apprêtent à les utiliser en 2017. Eurostat encourage très largement le recours à ce type de données. Des recommandations sur la manière d'intégrer les données de caisse dans le calcul de l'indice sont par ailleurs en cours de rédaction sous l'égide d'Eurostat.

2 - Modalités de la collecte des données de caisse

2.1 – L'utilisation des données de caisse dans l'IPC couvrirait 1/6e de la consommation des ménages

L'indice des prix à la consommation couvre la totalité de la consommation de biens et services marchands sur l'ensemble du territoire. Toutes les formes de vente y sont suivies : supermarchés, hypermarchés, petits magasins traditionnels, magasin populaire, marché, internet, services... Toutes les formes de vente et tous les types de produits ne se prêtent pas naturellement à un suivi des prix par les données de caisse ; les prix pratiqués sur les marchés et par les artisans (boucherie, boulangerie, …) ne peuvent pas être collectés ainsi ; la plupart de leurs produits ne sont pas en effet munis de codes-barres et il n'est pas envisageable de leur demander un transfert quotidien de leurs données de caisse. La collecte traditionnelle par enquêteur sera donc pour partie maintenue.

Les entreprises dont les données de caisse seront recueillies par l'Insee sont classées dans la NAF rév. 2, 2008, édition 2015 dans la classe 47.11 « Commerce de détail en magasin non spécialisé à prédominance alimentaire ». Au sein de cette classe, seules les sous-classes :

- 47.11D – supermarchés ;
- 47.11E – magasins multi-commerces ;
- 47.11F – hypermarchés ;

entrent actuellement dans le champ de l'enquête.

D'ici quelques années, ce dispositif de collecte pourrait être étendu à d'autres formes de ventes (supérettes, grandes surfaces non alimentaires) et à d'autres secteurs d'activité (articles électroménagers, de bricolage, de jardinerie, d'ameublement, d'habillement …).

Le choix de limiter, dans un premier temps, le champ des entreprises couvertes tient au mode de collecte des données (voir point 2.2) afin de contenir tout à la fois la charge de fourniture des données pour les entreprises et la charge de traitement de ces données pour l'Insee (nombre d'interlocuteurs et de transmission à mettre en place, nombre de formats de fichiers différents). Il tient également à la difficulté de traiter les données de caisse pour certains types de produits.

En effet, l'utilisation des données de caisse sera, dans un premier temps, circonscrite aux articles de grande consommation (articles alimentaires, hors produits frais, article d'entretien et d'hygiène-beauté) vendus dans la grande distribution alimentaire. Les produits frais (boulangerie au rayon frais, viande à la découpe, fruits et légumes, etc.) sont exclus en raison de l'absence d'EAN[2], identifiant commun à l'ensemble des enseignes, attribué par le producteur. Les autres articles vendus dans la grande distribution (vêtements, électroménager, etc.) ont été exclus en raison d'une complexité méthodologique de traitement plus importante ; de même pour les articles des commerces spécialisés en équipement du foyer, biens culturels, etc. En particulier, ces produits sont rapidement remplacés (nouvelles collections, innovations technologiques) et il est plus complexe de neutraliser lors des remplacements les modifications de qualité (voir point 1.2 pour le traitement de l'effet qualité lors des remplacements). L'extension de l'utilisation des données de caisse à ces secteurs est cependant envisagée à terme.

Le champ des données de caisse, dans sa première version restreinte en termes de commerces et de produits suivis, représente de l'ordre de 17 % de la consommation des ménages[3].

Le champ géographique de l'enquête est la France métropolitaine.

2.2 – Des modalités de collecte conçues pour minimiser la charge pour les enseignes et garantir la sécurité et la confidentialité des données.

Les fichiers que l'Insee souhaite collecter sont ceux que la plupart des enseignes transmettent déjà quotidiennement à deux sociétés d'études de marché, IRI Wordwide et Nielsen France. Ils indiquent, pour chaque article vendu un jour donné dans l'un des établissements de l'enseigne :

- le numéro EAN permettant d'identifier l'article ;
- l'identifiant du point de vente tel que fournit aux sociétés IRI Wordwide et Nielsen France ;
- la date des ventes ;
- le nombre d'articles vendus au cours de la journée ;
- le prix unitaire de l'article pour cette journée ;
- et/ou le chiffre d'affaires généré par les ventes de l'article durant la journée ;
- le ou les libellés descriptifs de l'article ;
- le code de classement de l'article dans la nomenclature interne de l'enseigne ;

et plus généralement, lorsqu'elles sont transmises quotidiennement par l'enseigne signataire aux sociétés prestataires citées ci-dessus pour les études et panels fondés sur les données de caisse, toutes les données complémentaires qui permettent d'identifier ou de classer l'article.

La collecte est exhaustive. Les enseignes transmettent les données relatives aux ventes (quantités vendues, prix et/ou chiffres d'affaires) de tous les articles (codes-barres) vendus dans chacun de leurs points de vente chaque jour. L'exhaustivité de la collecte permet à l'Insee de produire des indices sans biais statistique en incluant dans le calcul toutes les données ou de constituer la base de sondage.

L'Insee collectera quotidiennement ces données - pour les enseignes qui transmettent leurs données de caisse aux sociétés citées ci-dessus - auprès de la société à laquelle elle achète les référentiels d'articles et de points de vente et qui fait office de tiers de transmission. Pour ces entreprises, la charge de transmission des données à l'Insee devrait donc être nulle. Actuellement, dans le cadre de

[2] European Article Number

[3] Source : note méthodologique relative aux indices mensuels des prix dans la grande distribution.

l'expérimentation menée avec les enseignes volontaires (voir 4.1), le tiers de transmission est la société IRI Worlwide. Le marché arrivant à son terme en novembre 2016, le tiers de transmission sera la société retenue dans le nouveau marché. Le nom de la société retenue sera communiqué aux enseignes dès la notification du marché.

Pour les enseignes qui ne transmettent pas actuellement leurs données à des tiers de transmission, la collecte se fera directement. Les modalités de cette récupération de données doivent être précisées avec les enseignes concernées, de manière à ce que des transmissions sécurisées soient assurées, comme elles le sont avec IRI Worldwide. Les caractéristiques données ci-dessous prévaudront également.

Les transmissions de données sont sécurisées grâce :

- au chiffrement des fichiers au format Opengpg et d'un échange de clés asymétriques à deux niveaux ;
- à l'établissement d'une communication sur la base d'un protocole FTP ou Pesit (CFT) sécurisée par un tunnel SSL.

Le stockage des données est réalisé dans des bases sécurisées auxquelles l'accès est nominatif. L'Insee s'y est engagé, fort de son expérience dans le traitement et la transmission de données (recensement, données fiscales, données de salaire, etc.).

La collecte, quotidienne, est prévue à partir du 1er janvier 2017.

Les données ainsi collectées seront conservées par l'Insee pendant au plus 3 années.

Une convention signée avec chacune des enseignes enquêtée reprendra les conditions techniques de l'accès aux données.

2.3 – La division des prix à la consommation de l'Insee assure la collecte des données

Au sein de l'Insee, le service de l'Insee responsable de la collecte des données de caisse est la division des prix à la consommation. Cette division de la direction générale de l'Insee appartient à la direction des statistiques démographiques et sociales et est située plus précisément au sein de l'unité des prix à la consommation et des enquêtes ménages. Elle est chargée de la conception, de la production, de la diffusion et de l'analyse de l'indice des prix à la consommation et des indices associés (grande distribution, indice sous-jacent, etc.). Elle pilote le système d'information sur les prix à la consommation. Elle réalise les études méthodologiques et thématiques sur l'indice des prix à la consommation. Elle participe à l'harmonisation internationale en matière d'indice de prix, en particulier dans le cadre européen. Elle produit la partie française de l'IPC harmonisé au niveau européen, en accord avec les directives communautaires. Composée d'une vingtaine de personnes, cette division a autorité sur les 9 sites prix de métropole et des DOM qui gèrent la collecte en magasins et sur un pôle spécialisé à Bordeaux qui traite une partie des relevés centralisés.

2.4 – L'Insee se propose de mettre en place une instance de dialogue avec les enseignes

La loi numérique prévoit l'utilisation d'un nouveau type de données, les données privées, à des fins de statistique publique. Pour les données de caisse, l'Insee propose la mise en place d'une instance de dialogue. Celle-ci se réunira au moins une fois par an. Elle permettra des échanges sur les difficultés rencontrées dans la transmission des données et de s'informer réciproquement des avancées sur l'utilisation des données par l'Insee ou des évolutions prévues par les enseignes. Elle

sera composée de représentants des enseignes, et des représentants de l'Insee en charge de la récupération et de l'utilisation des données, informaticiens et méthodologues de la division des prix de l'Insee.

3 - Exploitation et diffusion des résultats

Les données de caisse serviront à la production :

- de l'indice des prix à la consommation ;
- d'indices détaillés par région et/ou par fonction de consommation de la COICOP[4] ;
- de prix moyens ;
- de comparaisons spatiales de prix.

Dans leur version harmonisée au niveau européen, les indices seront intégrés au calcul de l'indice des prix à la consommation harmonisé.

Les indices de prix agrégés seront diffusés par l'Insee sur son site internet (insee.fr) et par Eurostat (ec.europa.eu/eurostat/fr/home).

Comme toutes les statistiques publiées par l'Insee, les indices et prix moyens diffusés seront couverts par le secret statistique. L'enquête sur les données de caisse fera partie des enquêtes validées par le Cnis[5] et relèvera à ce titre de la loi n°51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques.

- toutes les personnes ayant accès aux données collectées (enquêteurs, statisticiens, chercheurs autorisés) sont astreintes au secret statistique (loi n° 51-711 du 7 juin 1951 ; article L226-13 du code pénal) ;
- les renseignements individuels d'ordre économique et financier communiqués au titre de cette enquête ne peuvent en aucun cas être utilisés à des fins de contrôle fiscal ou de répression économique (loi du 7 juin 1951) ;
- selon les règles en vigueur, les données publiées à partir des enquêtes et du recensement de la population ne permettent une identification ni directe ni indirecte des répondants et de leurs réponses. Pour les données relatives aux entreprises, l'Insee ne publie aucun résultat qui concerne moins de trois entreprises ou établissements. De même, un résultat ne sera diffusé que si aucune entreprise ou établissement ne contribue à plus de 85 % de ce résultat.

La loi et la pratique françaises s'inscrivent dans un cadre européen homogène énoncé dans le « code de bonnes pratiques de la statistique européenne » (http://ec.europa.eu/eurostat/fr/web/quality/european-statistics-code-of-practice).

4 - Étude de faisabilité

4.1 – Une expérimentation menée depuis plusieurs années...

Une expérimentation de calcul d'indice à partir des données de caisse est menée depuis 2012. Quatre enseignes de la grande distribution alimentaire, représentant environ 30% du marché, y participent sur la base du volontariat. La transmission des données y est encadrée par des

[4] La classification des fonctions de consommation des ménages (Classification of Individual Consumption by Purpose - COICOP) est une nomenclature permettant de décomposer la consommation des ménages par unités de besoin.

[5] Conseil National de l'Information Statistique

conventions signées avec chaque enseigne. Les données transmises se rapportent aux ventes quotidiennes de chaque article (identifié par son code-barres) dans chacun des points de vente de l'enseigne en France métropolitaine (hors Corse). Elles comprennent, pour chaque article vendu dans un point de vente un jour donné, la quantité d'articles vendue, le prix de vente (et/ou le chiffre d'affaires généré), un court descriptif de l'article (semblable à celui qui figure sur les tickets de caisse) ainsi que l'identifiant de la famille dans laquelle l'enseigne classe l'article.

La mise en place de cette enquête expérimentale a fait suite à une étude de faisabilité menée en 2011 à partir d'un jeu unique de données agrégées à la semaine pour dix familles de produits (yaourts, tablettes de chocolat, …) et un échantillon de mille points de vente d'enseignes volontaires. Les données portaient sur les trois années 2007, 2008 et 2009.

4.2 - … qui a permis d'identifier et de traiter les problèmes statistiques liés à l'utilisation des données de caisse ...

L'expérimentation a permis le traitement approfondi de différentes questions statistiques.

L'exploitation des données a permis de calculer un indice expérimental selon la méthodologie du futur indice issu des données de caisse. Cet indice expérimental a été calculé pour l'année 2014, avec comme année de base 2013, sur le champ de l'alimentaire (fonction 01 de la COICOP - produits alimentaires et boissons non alcoolisées). Les résultats montrent une bonne proximité des indices issus des données de caisse avec les indices issus de la collecte terrain au niveau 3 de la COICOP (par exemple, classe 01.2.1 - café, thé et cacao). Les différences entre les indices sont principalement dues à la précision des indices issus de la collecte terrain à ce niveau de détail. Certains indices divergent plus fortement et ce pour diverses raisons. Les indices de la viande (classe 01.1.2 de la COICOP) divergent en raison d'un « défaut de couverture » des données de caisse par rapport à la collecte en magasin qui est la référence dans cette comparaison. En effet, d'une part, la viande se vend encore beaucoup dans les boucheries traditionnelles et sur les marchés et, d'autre part, l'indice issu des données de caisse ne prend pas en compte les viandes vendues, dans la grande distribution, à la coupe (rayon boucherie traditionnelle) ou pré-conditionnées dans le magasin. Le nombre de relevés dans les rayons de boucherie traditionnelle des grandes surfaces est trop faible pour permettre de recalculer, avec précision, un indice à partir de la collecte en magasins sur un champ comparable à celui des données de caisse. Cette divergence sera résorbée en maintenant la collecte terrain dans un certain nombre de commerces traditionnels et dans les rayons boucherie des hypermarchés et supermarchés. Les indices des poissons et crustacés préparés en conserve, surgelés et fumés (sous-classe 01.1.3.2 de la COICOP) divergent, en début d'année, en raison des rabais importants sur les produits conditionnés pour les fêtes de fin d'année (saumon fumé par 10 ou 12 tranches par exemple). Ces produits, présents deux mois de l'année, ne sont actuellement pas suivis dans l'indice des prix à la consommation (IPC). Le recours aux données de caisse permet dans ce cas d'améliorer la qualité de l'indice en augmentant sa couverture et en suivant des phénomènes qui ne pouvaient pas être observés auparavant, faute de données.

L'Insee a choisi de calculer l'indice issu des données de caisse en conservant les concepts sous-jacents au calcul de l'IPC à partir de la collecte par enquêteurs. Pragmatiquement, l'Insee ne s'interdit pas de simplifier les traitements des données (qui peuvent s'avérer très coûteux sur les volumes considérables que représentent les données de caisse) s'il est prouvé qu'il n'y a pas de différence entre le traitement qui suit les concepts de l'IPC et un traitement alternatif plus simple. Dans cette méthodologie, l'indice mesure, chaque mois, le coût d'un panier de biens et services dont la composition est fixée en début d'année à partir des ventes de l'année précédente. La problématique des remplacements de produits en cours d'année a donc fait l'objet d'études approfondies. Une méthode automatique de choix des produits remplaçants a été mise au point. Les

formules d'ajustement de la qualité, nécessaire lors d'un remplacement de produit, ont été testées puisque les données de caisse, qui apportent une connaissance de l'historique des prix du produit remplaçant, offrent l'opportunité d'utiliser des méthodes impossibles à mettre en œuvre dans la collecte classique. Ces travaux ont fait l'objet de multiples présentations dans des séminaires internes à l'Insee et internationaux (workshop européens, groupe d'Ottawa).

Les données de caisse, en limitant le coût de collecte de données supplémentaires, permettent de relever les prix de plusieurs articles dans une même variété (sous-groupe de la COICOP-5) et un même point de vente. Le niveau d'agrégation des données de caisse a été redéfini. Une réflexion théorique et pratique a été menée sur le choix de l'indice permettant d'agréger ces données entre elles en respectant les hypothèses de substitutions réalisées par le consommateur et sur la manière d'agréger ces indices de manière à obtenir des indices de postes (COICOP-5) qui pourront être agrégés eux-mêmes aux indices de même niveau calculés à partir de la collecte terrain. Ces travaux ont été présentés dans les mêmes instances que ceux sur les remplacements de produits.

L'exploration des données quotidiennes a également permis de mieux comprendre l'intégration des rabais et remises à l'initiative des enseignes dans les données et de mettre au point une méthode de prise en compte des promotions à l'initiative des fabricants qui, en changeant le conditionnement de l'article, changent son identifiant (code-barres). Cette exploration a fait émerger des problématiques propres aux données de caisse telles la saisonnalité des produits qui doit être envisagée par produit et non, comme dans la collecte classique, au niveau d'une variété. Les saumons fumés conditionnés par 10 ou 12 tranches pour les fêtes de fin d'année en sont un exemple.

Enfin, des essais de comparaisons spatiales des prix, fondées sur un modèle économétrique dans lequel le niveau de prix est expliqué par un effet fixe spatial, conditionnellement à un effet fixe code-barres, ont été menés. Cette méthode permettrait de réaliser des comparaisons spatiales de prix - très coûteuses car fondée sur une enquête terrain - plus fréquemment.

En sus de ces études approfondies sur des points précis qui ont fait l'objet de nombreuses présentations, l'architecture du futur indice des prix utilisant les données de caisse a été pensée : une application, en cours de développement, permettra de recevoir les fichiers de transmission quotidiens, de les contrôler puis de traiter les observations (le classement des données selon une nomenclature de fonction de consommation est notamment un enjeu), de calculer des indices élémentaires respectant la méthodologie de l'indice des prix (panier fixe, traitement des promotions, de la saisonnalité, des remplacements, des substitutions…) puis d'intégrer ces informations dans le calcul global de l'indice des prix à la consommation.

4.3 - … de démontrer également la capacité informatique de l'Insee à traiter ces données ...

Côté informatique, le traitement des données de caisse soulève des problématiques liées à la réception quotidienne d'un volume important de données (de l'ordre de 10 Go par semaine). L'expérimentation a permis d'estimer la volumétrie globale des données reçues à terme, ce qui a permis d'imaginer une architecture informatique adaptée aux contraintes portant sur le futur système d'information.

Par ailleurs, dans le cadre d'un calcul mensuel d'indices avec des contraintes très fortes sur les délais (la première estimation de l'indice des prix à la consommation d'un mois donné est publiée dès la fin de ce mois), une très forte attention a été portée à la capacité à prendre en compte très rapidement de potentiels changements dans les données en entrée (format des fichiers, contenus).

4.4 - … et d'évaluer le coût financier de l'opération

Côté Insee, au sein de la division des prix à la consommation, une équipe de trois personnes sera chargée du traitement statistique des données de caisse. Ce traitement ne peut être réalisé sans l'appui des services informatiques de l'Insee, que ce soit pour la réception des fichiers qui arriveront chiffrés et compressés, pour l'exécution des traitements et pour la maintenance évolutive des programmes et applications. Le traitement des données de caisse permettra néanmoins à l'institut de réaliser des économies sur la collecte en magasins effectuée jusque là par des enquêteurs, et de consacrer une partie plus importante de ses forces d'enquêtes à ses autres besoins.

Pour les enseignes qui transmettent déjà leurs données à une société tiers de transmission, la transmission des données à l'Insee pourra se faire *via* cette société, a priori sans coût supplémentaire. Pour les enseignes qui ne le font pas, le coût de cette transmission spécifique de données à l'Insee est à l'étude.

4.5 La concertation avec les enseignes

Au delà des enseignes participant à l'expérimentation, l'Insee a souhaité recevoir l'avis des enseignes de la grande distribution sur la faisabilité de la transmission des données de caisse. Une réunion a été organisée à cet effet à l'Insee le 30 juin 2016. Seules deux enseignes n'étaient pas présentes.

Cette réunion a permis de présenter le projet dans le détail, mais surtout les modalités techniques pressenties pour la transmission des données. De premiers échanges ont eu lieu à cette occasion. Les enseignes ont souligné leur attachement à pouvoir contractualiser avec l'Insee sur ce projet, malgré la dimension réglementaire. Elles ont exprimé le souhait d'être associées à la rédaction de l'arrêté en cours de préparation définissant les modalités précises des transmissions attendues par l'Insee.

Les enseignes ont également confirmé que la grande majorité d'entre elles transmet déjà les données de caisse à des sociétés tiers, sociétés d'étude de marché, qui peuvent assurer la transmission des données pour le compte des enseignes. Il sera donc possible pour l'Insee de récupérer les données via l'une de ces sociétés. Une enseigne a néanmoins exprimé le souhait de transmission directe des données à l'Insee, de manière à ne pas être liée. Une solution technique est à l'étude compte tenu des outils de transmission utilisés. Une enseigne s'est manifestée anticipant des difficultés dans la transmission de ses données au vu de son système d'information. Une expertise plus approfondie est en cours. De manière plus générale, cette réunion a permis des échanges sur les difficultés techniques des enseignes par rapport au projet. Celles-ci s'avèrent plutôt mineures de manière générale, et des solutions devraient pouvoir être trouvées.