UTILISER LES DONNÉES DE CAISSES POUR LE CALCUL DE L'INDICE DES PRIX À LA CONSOMMATION

*Marie Leclair**

Les données de caisses sont des données recueillies par les enseignes de la grande distribution quand le consommateur passe à la caisse des magasins. Ces données très volumineuses seront utilisées pour calculer l'indice des prix à la consommation (IPC) à partir de janvier 2020. La couverture exhaustive du champ par ces données, la connaissance fine de la consommation des ménages qu'elles apportent sont des avancées majeures pour la production de statistiques des prix : elles en améliorent, notamment, la précision et permettent à terme de produire de nouvelles statistiques (indices plus détaillés, indices régionaux, prix moyens, comparaisons spatiales des prix). Cependant, elles obligent également à trouver un certain nombre de solutions nouvelles, en particulier en termes d'automatisation des traitements. Le choix d'une solution informatique big data, le recours à un référentiel des codes-barres contribuent à pouvoir exploiter ces données, tout en conservant les concepts de l'IPC, notamment le recours à un panier fixe de produits.

Scanner data are data gathered by large retailers when consumers go to pay for their goods in store. From January 2020 onwards, these enormous volumes of data will be used to calculate the consumer price index (CPI). The comprehensive coverage of the field by these data and the detailed knowledge of household consumption that they provide are major steps forward when it comes to producing price statistics: they improve, in particular, their accuracy and ultimately mean that new statistics can be produced (more detailed indexes, regional indexes, average prices and spatial price comparisons). However, they also require that a number of new solutions be found, in particular in terms of automated processing. Choosing a big data IT solution and using a barcode dictionary help enable these data to be exploited whilst preserving the concepts of the CPI, in particular the reference to a fixed basket of goods.

L'indice des prix à la consommation (IPC) français est calculé à partir de 200 000 relevés mensuels effectués par des enquêteurs dans des points de vente physiques. Cette collecte de terrain a été progressivement complétée par d'autres sources, dématérialisées : relevés sur internet et données administratives totalisent 190 000 prix additionnels chaque mois. À partir de janvier 2020, une nouvelle source, d'une toute autre ampleur sera utilisée : les données de caisses.

*Cheffe de la division des Prix à la consommation, Insee,
marie.leclair@insee.fr

UNE NOUVELLE SOURCE DE DONNÉES, LES DONNÉES PRIVÉES

Les données de caisses sont les informations sur les prix payés et les produits achetés, recueillies par les enseignes au moment où le consommateur passe à la caisse des magasins.

> *« Les données de caisses sont les informations sur les prix payés et les produits achetés, recueillies par les enseignes au moment où le consommateur passe à la caisse des magasins. »*

Beaucoup plus volumineuses que les données mobilisées jusqu'à présent pour le calcul de l'IPC (1,7 milliard d'enregistrements reçus chaque mois), ces données privées sont une réelle opportunité pour le calcul de statistiques de prix. Mais ces données privées posent également de nouvelles questions, sur leur accès tout d'abord, sur leur fiabilité ensuite, et enfin sur la capacité de l'Insee à les exploiter à des fins statistiques et d'un point de vue informatique.

Les données de caisses, qu'on appelle plus génériquement données de transaction, sont disponibles pour bien des magasins, dès lors que les transactions sont enregistrées. Si la centralisation en favorise la mise à disposition, le traitement statistique de ces fichiers n'est pas pour autant aisé pour tous les produits de la consommation : à ce stade, l'Insee n'exploite « que » les données des enseignes de la grande distribution (super et hypermarchés) de France métropolitaine, pour les produits alimentaires industriels, d'entretien et d'hygiène-beauté.

UNE OPPORTUNITÉ POUR LE CALCUL DE STATISTIQUES DE PRIX

Les données de caisses sont mobilisées depuis de nombreuses années par des panélistes à des fins d'études de marché. Certains instituts nationaux statistiques ont vu assez tôt l'intérêt de mobiliser de telles sources pour le calcul de leur IPC : les Pays-Bas les utilisent ainsi depuis 2002, suivis par la Norvège en 2005, la Suisse (2008), la Suède (2012), la Belgique (2015), le Danemark (2016), l'Islande (2016), le Luxembourg et l'Italie (2018). Eurostat a contribué à l'extension de leur utilisation, *via* des *grants,* des *workshops* et un manuel (Eurostat, 2017). L'utilisation des données de caisses est également un sujet récurrent du groupe des experts prix de l'UNECE[1] et du BIT[2], réunissant des universitaires et des statisticiens du monde entier (UNECE, 2018).

Différentes motivations expliquent cet intérêt : le souhait de faire entrer les instituts statistiques dans l'ère du big data, le recours à des données privées produites « gratuitement » (même si leur exploitation peut être coûteuse), l'exhaustivité des données mais également l'apport de nouvelles informations sur les produits consommés, non disponibles jusqu'à présent et qui ouvrent de nombreuses opportunités pour les statistiques de prix, comme on le verra par la suite.

1. Commission économique pour l'Europe des Nations unies (CEE-ONU), en anglais *United Nations Economic Commission for Europe* (UNECE).
2. Bureau international du travail.

Dans le cas français, les données de caisses sont des données exhaustives sur leur champ ; elles sont collectées en continu et centralisées selon un pas journalier. Ainsi pour chaque code-barres de produits[3] ***(figure 1)***, sont enregistrés les quantités vendues du produit ainsi que le prix ou le chiffre d'affaires associé, par points de vente et jour de vente ***(figure 2)***.

DES STATISTIQUES PLUS PRÉCISES ET PLUS DÉTAILLÉES

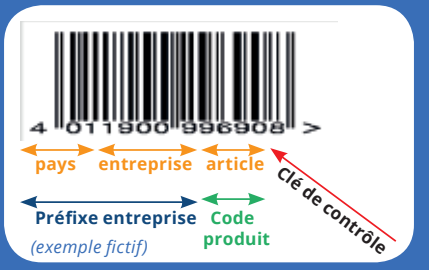
Selon Tassi (2019), les données *big data* se caractérisent par la quantité d'information (pouvant aller jusqu'à l'exhaustivité sur un domaine donné) et la fréquence d'acquisition de cette information. Dans le cas des statistiques de prix, la disponibilité des données selon un pas quotidien et en continu[4] est importante pour produire un indice mensuel comme l'IPC, qui plus est selon des contraintes temporelles fortes[5]. Il n'est cependant pas prévu de produire un IPC infra-mensuel.

L'exhaustivité de la source permet, en revanche, de produire des statistiques plus précises et d'envisager de produire des indices plus détaillés, par exemple sur des segments de consommation spécifiques. L'IPC est déjà produit mensuellement pour plus

Figure 1. La structure d'un code-barres

Tout article destiné à la vente au consommateur dans un magasin de détail doit être identifié avec un code à barres. Le code-barres est la transcription graphique d'un code GTIN *(Global Trade Item Number)*. Le code GTIN et les barres forment un tout indissociable pour sa reconnaissance en caisse. Les barres permettent la saisie automatique des données à chaque lecture de l'article.

Un code-barres est donc *un identifiant unique de produit, respectant une clé et donnant une information sur le pays et l'entreprise de production, mais il ne permet pas de donner une information sur la nature du produit suivi.*



Le GTIN (nom actuel et officiel), l'EAN (ancienne dénomination Européenne) ou le GENCOD (ancienne dénomination française), désignent le code qui identifie une unité commerciale c'est-à-dire une unité dont le prix peut être fixé, qui peut être commandée, livrée ou facturée aux fins d'échanges commerciaux. Ces unités commerciales peuvent être des unités de vente au consommateur, des unités logistiques (cartons, palettes, box, bacs…), des unités d'expédition. Le code GTIN est inscrit le plus souvent en chiffres au-dessous des barres. Il en existe différents formats, dont le plus fréquemment utilisé est le GTIN 13, soit 13 chiffres structurés de la manière suivante :

Préfixe entreprise : 6, 7, 8, 9, 10 ou 11 chiffres selon les besoins de codification de l'entreprise.

Code produit : attribué par le propriétaire de la marque commerciale qui dispose de 6, 5, 4, 3, 2 ou 1 chiffres selon la longueur du préfixe entreprise. C'est la seule partie variable du code, incrémentée par l'entreprise.

Clé de contrôle : 1 caractère, calculé sur la base des chiffres précédents et permet d'éviter toute erreur de saisie.

3. Appelé également GTIN pour *Global Trade Item Number* ou EAN pour *European Article Numbering*.
4. Les données sont envoyées à l'Insee avec deux jours de décalage.
5. Une estimation provisoire de l'IPC est publiée dès le dernier jour ouvré du mois.

de 250 sous-classes et annuellement pour plus de 360 postes[6]. Une demande de la part du public d'information toujours plus détaillée existe, mais ce n'est pas forcément le principal apport des données de caisses.

C'est la dimension géographique de l'exhaustivité des données de caisses qui est particulièrement intéressante. Jusqu'à présent, les prix sont relevés par l'IPC dans un échantillon de communes de plus de 2 000 habitants, représentatif au niveau national. Or l'Insee rencontre une difficulté pratique dans les plus petites communes car, du fait d'un tissu commercial moins dense, les enquêteurs doivent parcourir beaucoup plus de kilomètres pour relever des prix.

Outre la fiabilisation de la collecte dans les zones plus rurales, l'exhaustivité permet d'être représentatif au niveau de chaque région et d'envisager de produire, à terme, des indices de prix régionaux (sur le champ des données de caisses) : jusqu'à présent, seuls un indice métropolitain et pour chaque département d'Outre-mer étaient publiés. Par ailleurs, la représentativité par territoire fin et le détail de l'information sur les produits suivis permettent d'envisager à terme des comparaisons spatiales de niveau de prix : cet exercice n'est réalisé actuellement que tous les 5-6 ans par l'Insee et, en France métropolitaine, uniquement entre l'agglomération parisienne, la Corse, et la province (Clé et *alii,* 2016). Des travaux expérimentaux (Léonard et *alii,* 2019) montrent que ces données de caisses peuvent être avantageusement mobilisées pour des comparaisons spatiales de prix. Elles sont d'ailleurs déjà utilisées par certains pays pour des comparaisons européennes de niveaux de prix, les parités de pouvoir d'achat.

Figure 2. Un échantillon des données de caisses

Identifiant du point de vente	EAN	Description de l'article	Date des ventes	Quantités vendues	Prix de vente (en €)	Chiffre d'affaires (en €)
723	3275770004817	[illegible] 150G	20140108	10	1.89	18.90
723	3155230040286	[illegible] BACON 150G	20140108	7	2.38	16.66
986	3185670001080	[illegible] STRAINED SOFT 6%MG 1KG	20140128	25	2.59	64.75

Permet de faire le lien avec le référentiel des points de vente
(localisation, surface)

Permet de faire le lien avec le référentiel des articles
(marque, description détaillée, volume)

6. L'indice des prix à la consommation est publié selon la nomenclature Coicop *(Classification of Individual Consumption by Purpose)* ; la sous-classe et le poste sont les deux derniers niveaux de cette nomenclature.

CONNAÎTRE LES QUANTITÉS CONSOMMÉES ET DISPOSER D'UNE BASE DE SONDAGE

Si les statisticiens des prix se sont autant intéressés aux données de caisses, c'est aussi parce qu'elles leur donnent accès à une information dont ils avaient jusqu'à présent une connaissance très grossière, ancienne et agrégée : le type de produits consommés par les ménages et les quantités consommées par ceux-ci.

Par exemple, la comptabilité nationale donne aujourd'hui une information sur le poids des céréales pour petit déjeuner dans la consommation sur le territoire français. Mais elle ne la détaille pas par variété de céréales ou encore moins par marques ou par points de vente.

On ne peut donc pas s'appuyer sur une base de sondage pour tirer un échantillon aléatoire de produits dont on suit le prix. Faute d'information, le panier de l'IPC est aujourd'hui *de facto* défini par une méthode de quotas : les unités urbaines dans lesquelles se rend l'enquêteur sont certes tirées aléatoirement en fonction de la population qui y vit et des habitudes de consommation qu'on lui impute (Jaluzot et Sillard, 2016). Mais le choix des points de vente précis et des produits suivis est effectué par l'enquêteur en fonction de quelques contraintes ou quotas (forme de vente, variétés de produits suivies). Dans le cas des céréales de petit déjeuner, on demandera à un enquêteur de se rendre dans une unité urbaine donnée, tirée aléatoirement, et d'y trouver par exemple 4 paquets de céréales de type muesli, dont un en hypermarché, deux en supermarché et un en supérette. C'est l'enquêteur qui choisira les points de vente, et dans ceux-ci la boîte de céréales qu'il retiendra.

« Avec les données de caisses, on dispose enfin d'une base de sondage : l'ensemble des articles vendus par points de vente, avec le poids de chacun de ces articles dans le chiffre d'affaires des points de vente. »

Avec les données de caisses, on dispose enfin d'une base de sondage : l'ensemble des articles vendus par points de vente, avec le poids de chacun de ces articles dans le chiffre d'affaires des points de vente. L'existence d'une base de sondage permet de recourir à un tirage aléatoire des produits d'une part, et d'autre part de maîtriser le biais d'échantillonnage. Elle permet également de repérer rapidement les produits nouveaux à introduire dans le panier IPC ou les produits en perte de vitesse qu'il faut retirer, afin que le panier soit toujours à jour et représentatif de la consommation des ménages.

Au final, à l'Insee, compte tenu des possibilités informatiques et d'automatisation (voir *infra*), c'est l'exhaustivité des données de caisses qui a été retenue, sans recourir à un échantillonnage.

MIEUX TRAITER LES SUBSTITUTIONS DU CONSOMMATEUR ENTRE LES PRODUITS

La connaissance des quantités précises vendues pour chaque article permet également de progresser dans la pratique des indices : pour calculer un indice synthétique des prix, les relevés sont agrégés *via* un certain nombre de formules ***(encadré)***. La théorie des indices définit les propriétés de ces formules et les indices qu'il convient de retenir d'un point de vue théorique (Sillard, 2016 ; FMI, 2004). Dans la pratique, l'absence de connaissance sur les quantités courantes consommées par les consommateurs à un niveau détaillé contraint le choix.

Pour l'IPC, on utilise :

- au niveau le plus agrégé (par exemple pour agréger les boîtes de céréales avec les pâtes alimentaires) un indice de type Laspeyres, avec une pondération utilisant les dépenses passées de consommation ;
- au niveau le plus fin (par exemple pour agréger les différentes boîtes de céréales entre elles), en l'absence d'information sur les quantités consommées, même passées, un indice de Dutot ou Jevons, impliquant une équipondération des prix relevés.

Par le détail de l'information qu'elles apportent, les données de caisses permettent de choisir des formules d'indice qui tiennent mieux compte au niveau le plus fin des substitutions effectuées par le consommateur entre deux produits (Leclair et *alii,* 2019) : lorsque le prix d'un produit augmente, l'impact sur l'utilité du consommateur peut être plus ou moins fort selon qu'il peut ou non reporter sa consommation vers un autre produit, plus ou moins substituable au premier. La capacité de traiter ces substitutions avait fait l'objet de débats, dans les années quatre-vingt-dix, sur l'éventualité d'un biais, à la baisse, des indices de prix qui les auraient insuffisamment prises en compte (Boskin, 1996). À l'époque, le recours aux données de caisses et à l'information fine sur les quantités vendues, donc

Encadré. Comment produit-on l'IPC ?

L'IPC mesure l'évolution des prix des produits consommés par les ménages. Les prix d'un panier fixe de produits sont suivis chaque mois de manière à mesurer une évolution pure de prix, à qualité constante. L'indice est un indice de type Laspeyres*, les différentes variétés de produits sont pondérées par leur poids passé dans la consommation des ménages. À un niveau plus fin que la variété des produits, les pondérations ne sont plus connues et des hypothèses sont effectuées pour agréger les prix élémentaires : les formules de Dutot** et de Jevons*** sont utilisées par l'IPC.

Afin de demeurer représentatifs de la consommation des ménages, les poids et le panier de produits suivis sont renouvelés chaque année : l'IPC est un indice chaîné annuellement. Par ailleurs, en cas de disparition d'un produit en cours d'année, celui-ci est remplacé par un produit proche et un ajustement qualité est effectué afin de corriger de l'écart de qualité entre le produit remplacé et remplaçant.

L'IPC est publié à un rythme mensuel, dès le dernier jour ouvré du mois pour l'indice provisoire, quinze jours environ après la fin du mois pour l'indice définitif. Cet indice définitif n'est par la suite plus révisé. Ces délais très courts et l'absence de révision imposent des contraintes très fortes au processus de production de l'IPC.

Outre les données de transaction, l'IPC utilise deux types de sources : des relevés de prix effectués par des enquêteurs de l'Insee sur le terrain (avant l'utilisation des données de caisses, de l'ordre de 200 000 relevés chaque mois dans des unités urbaines représentatives du territoire français) dans diverses formes de vente (y compris internet) ; des relevés collectés de manière centralisée soit que le prix de ces produits soit unique sur tout le territoire (service de télécommunication, électricité, tabac…), soit que des bases de données puissent être mobilisées pour calculer les évolutions de prix (données de la Caisse nationale de l'assurance maladie pour les services de santé, par exemple).

* L'indice de Laspeyres est un indice à panier fixe rapportant la moyenne des prix de la période courante à la moyenne des prix de la période de référence, en pondérant les prix courants et de la période de référence par les quantités consommées au cours de la période de référence.
** Un indice de Dutot est un indice à panier fixe rapportant la moyenne arithmétique des prix de la période courante à la moyenne arithmétique des prix de la période de référence. Tous les prix du panier sont équipondérés.
*** Un indice de Jevons est un indice à panier fixe rapportant la moyenne géométrique des prix de la période courante à la moyenne géométrique des prix de la période de référence. Tous les prix du panier sont équipondérés.

aux ajustements effectués par les consommateurs sur leur panier, avait déjà été présenté comme une solution prometteuse (Lequiller, 1997).

RÉPONDRE À DES POLÉMIQUES RÉCURRENTES AUTOUR DE L'IPC

Les données de caisses permettent également d'apporter des éléments de réponse au débat sur une sous-estimation possible de l'inflation par l'IPC : après le passage à l'euro, l'écart entre l'inflation ressentie par les ménages et l'inflation mesurée par l'Insee s'est accru (Leclair et Passeron, 2017). Une des explications de cet écart est que les ménages ont tendance à ne pas considérer certaines améliorations de la qualité des produits qu'ils consomment, alors que l'IPC les neutralise : une amélioration de la qualité d'un produit à prix inchangé se traduit dans l'IPC comme une baisse de prix. Or, si la norme sociale est modifiée vers des produits de meilleure qualité, et en conséquence plus onéreux, il est probable que le consommateur se sentira contraint à cette dépense de qualité supérieure et ressentira ce déplacement de la norme sociale comme une hausse du coût de la vie.

Une réponse possible était de construire des statistiques plus proches du ressenti, tout en continuant à produire l'IPC, qui est pertinent pour mesurer, notamment, la croissance du PIB en volume et le pouvoir d'achat des ménages. L'idée est de calculer des prix moyens d'ensemble de produits, qui prennent en compte les évolutions des habitudes de consommation et ne neutralisent pas ces changements de qualité (Moati et Rochefort, 2008). Par exemple, le développement de la consommation de riz de type basmati ou thaï, en général plus cher que du riz ordinaire, peut être considéré comme un changement de qualité du produit et ne se traduira pas par une augmentation de l'indice des prix du riz dans l'IPC. En revanche, le prix moyen du riz (toutes catégories confondues) augmentera du fait du poids croissant de ces riz parfumés dans les achats.

Les données de caisses permettent de calculer de tels prix moyens en donnant une information précise sur les quantités consommées, en plus des prix pratiqués.

UN MEILLEUR SUIVI DES PRIX EFFECTIVEMENT PAYÉS PAR LE CONSOMMATEUR

De manière peut-être un peu surprenante, un autre avantage des données de caisses est qu'elles permettent de mieux suivre le concept de prix que l'on souhaite mesurer avec l'IPC, par rapport à une enquête spécifique classique.

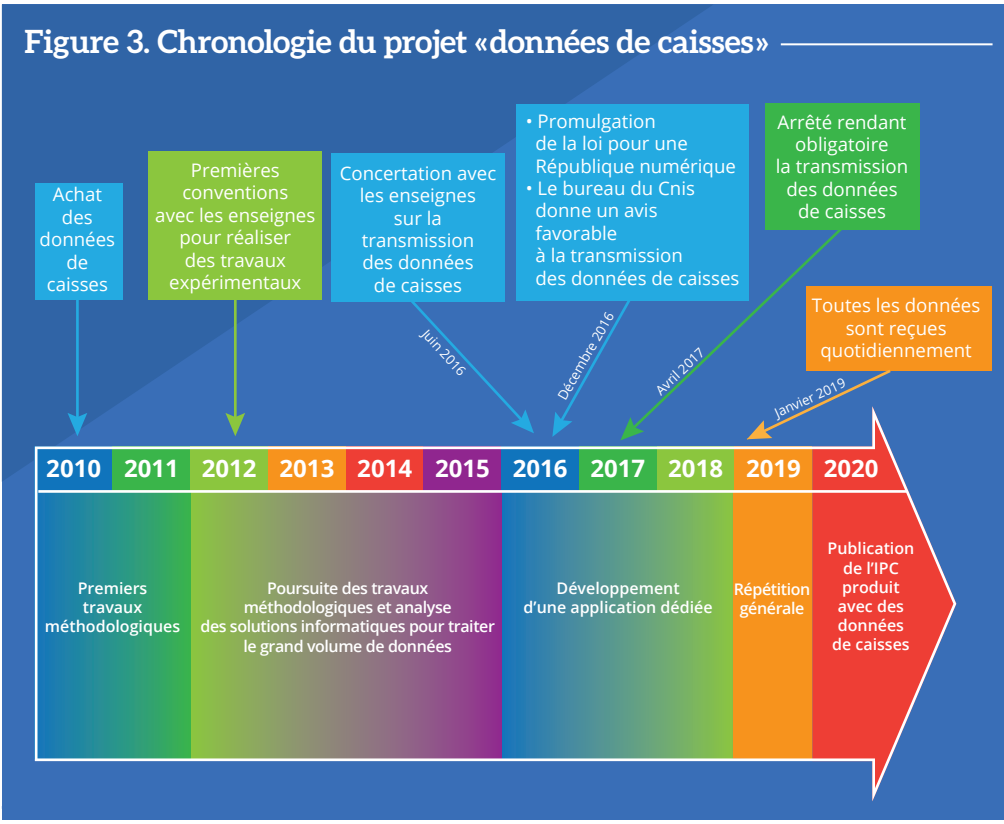
En effet, les enquêteurs envoyés par l'Insee sur le terrain ne peuvent relever que les prix affichés dans les points de vente. Or, ceux-ci peuvent diverger des prix réellement payés par le consommateur : ces différences s'expliquent en partie par des erreurs d'affichage, mais surtout par la mise en œuvre de certaines promotions. Actuellement, l'IPC ne mesure les promotions que lorsqu'elles s'appliquent à l'ensemble des consommateurs. Cette méthode, conforme à la réglementation européenne sur les indices de prix à la consommation, est la conséquence d'un manque d'informations sur le nombre d'acheteurs bénéficiant effectivement des promotions « discriminantes », par exemple liées au fait d'être porteur d'une carte du magasin.
Dans les données de caisses, ce sont les prix effectivement payés qui sont enregistrés. Elles intègrent ainsi de nombreuses promotions, même si certaines pratiques commerciales leur échappent encore, comme le « cagnottage » qui consiste à donner des points en vue d'une consommation future, en contrepartie de l'achat d'un produit spécifique.

LE LONG CHEMIN POUR ACCÉDER AUX DONNÉES PRIVÉES

Si l'apport des données de caisses pour la production de statistiques de prix est incontestable, leur utilisation par l'Insee n'est pas sans poser un certain nombre de difficultés.

La première, dans l'ordre chronologique, est tout simplement de pouvoir y accéder : les données de caisses sont des actifs incorporels des entreprises qui les produisent, et celles-ci n'avaient donc pas d'obligation à en donner un accès à l'Insee, même à des fins d'intérêt général pour la production de statistiques publiques. Dans un premier temps, des contacts ont été pris avec certaines enseignes afin de les convaincre de les transmettre à l'Insee. Depuis 2012, quatre enseignes (40 % environ du marché de la grande distribution) fournissaient ainsi les données à titre expérimental et dans le cadre de conventions ***(figure 3)***. Pour d'une part obtenir les données sur l'ensemble du champ des super et hypermarchés, et d'autre part pérenniser la mise à disposition, la loi de 1951 sur l'obligation, la coordination et le secret en matière statistique a été amendée. La loi[7] prévoit désormais la possibilité de rendre obligatoire la transmission de certaines données privées, après concertation des acteurs et uniquement pour remplacer des enquêtes statistiques obligatoires. Cet amendement, outre l'accès aux données de caisses, pourrait permettre de faciliter, à terme, l'accès à d'autres données privées.

Figure 3. Chronologie du projet «données de caisses»



7. Article 19 de la loi du 7 octobre 2016 pour une République numérique.

Après une concertation avec les enseignes de la grande distribution en juin 2016, une étude préalable de faisabilité et d'opportunité de l'utilisation des données de caisses pour l'IPC a été présentée fin 2016 au Conseil national de l'information statistique (Cnis). Après avoir reçu un avis favorable, un arrêté[8] a été signé par le ministre le 13 avril 2017, rendant obligatoire la transmission des données de caisses par les commerces de détail en magasin non spécialisé à prédominance alimentaire de plus de 400 m^2. Depuis janvier 2019, l'ensemble des données de caisses de la grande distribution alimentaire, hors hard discount, sont ainsi reçues quotidiennement par l'Insee.

DES STATISTIQUES À PARTIR DE DONNÉES PRODUITES À D'AUTRES FINS ?

Une deuxième difficulté dans l'utilisation des données de caisses est qu'elles n'ont pas été produites à des fins de production statistique. Cela pose deux questions : les statistiques produites à partir de ces données le sont-elles en toute indépendance et impartialité ? Les informations collectées sont-elles en adéquation avec l'objectif statistique recherché ?

Concernant le premier point, l'importance de l'IPC dans le débat public et l'existence jusqu'aux années soixante-dix de polémiques sur la manipulation de l'indice (Jany-Catrice, 2018) posent la question de la confiance que l'on peut avoir dans des données produites par des acteurs privés. Bien qu'il paraisse difficile que les enseignes puissent manipuler des données si volumineuses, l'Insee a souhaité garantir la qualité des données utilisées pour le calcul de l'IPC en organisant des enquêtes de contrôle. À l'avenir, chaque mois, un certain nombre de prix enregistrés dans les données de caisses seront contrôlés par des enquêteurs dans les points de vente. Dès 2019, une double collecte des prix, par les enquêteurs d'une part pour le calcul de l'IPC, dans les données de caisses d'autre part, permet de s'assurer de l'absence de divergence.

« Bien qu'il paraisse difficile que les enseignes puissent manipuler des données si volumineuses, l'Insee a souhaité garantir la qualité des données utilisées pour le calcul de l'IPC en organisant des enquêtes de contrôle. »

La seconde question porte sur le fait que ces données n'ont pas été produites initialement pour l'établissement de statistiques. Comme pour les données administratives (Rivière, 2018), il est possible que les données recueillies, les définitions retenues, le champ, ne correspondent pas exactement à ce que le statisticien souhaite mesurer. Ces faiblesses sont encore plus marquées pour le *big data* (Blanchet et Givord, 2017), caractérisé, non seulement par leur « Volume » et par la « Vélocité » de leur mise à disposition, mais également par un troisième « V », la Variété, soulignant l'aspect bien souvent peu structuré de ces données.

Dans ce paysage, les données de caisses semblent un cas un peu à part. Car pour commencer, l'information recueillie peut se révéler plus pertinente que dans le cas d'une enquête dédiée à la collecte des prix : le concept de prix est par exemple mieux suivi dans les données de caisses que par l'enquêteur qui ne peut collecter que le prix affiché ; l'information sur les quantités consommées est difficilement mesurable par une enquête, en tout cas, pas avec le

8. Arrêté du 13 avril 2017 rendant obligatoire la transmission des données par voie électronique à des fins de statistiques publiques.

degré de finesse nécessaire ; le champ des données de caisses peut être aisément complété par des données d'enquêtes (dans les autres formes de vente, pour les produits frais vendus en super et hypermarché, par exemple) pour couvrir l'ensemble de la consommation des ménages ; il couvre même dans certains cas mieux la consommation des ménages, en intégrant les données sur le *drive* par exemple, non couvert par l'IPC jusqu'à présent.

Par ailleurs, les données de caisses se distinguent au sein du *big data* parce qu'il s'agit en réalité de données très structurées, qui n'ont du *big data* que les deux premiers « V » du volume et de la vélocité.

Au total, alors que d'autres pistes d'utilisation du *big data* visent plutôt à produire de nouvelles statistiques, complémentaires mais non substituts de statistiques publiques existantes, les données de caisses ont cela de spécifique qu'elles peuvent réellement se substituer à des données d'enquête sans avoir à modifier les concepts ou le cadre méthodologique de ce que l'on veut mesurer.

1,7 MILLIARD D'ENREGISTREMENTS : NOUVELLE ARCHITECTURE INFORMATIQUE…

Si le choix de l'Insee est donc de considérer que les données de caisses peuvent remplacer les relevés effectués par les enquêteurs sans que cela nécessite d'adaptation des concepts utilisés pour l'IPC, la volumétrie des données à traiter nécessite un certain nombre de solutions, informatiques tout d'abord mais également statistiques ensuite, pour automatiser des traitements préalablement effectués manuellement.

L'Insee réceptionne chaque mois 1,7 milliard d'enregistrements, qui correspondent aux lignes dans les données de caisses, c'est-à-dire les ventes associées dans un point de vente pour un code-barres donné et un jour donné ***(figures 1 et 2)***. Autant d'enregistrements[9] à traiter mensuellement n'auraient pu être gérés par des bases de données classiques, dites relationnelles. Des technologies adaptées au *big data* (en l'occurrence le système Hadoop) ont été retenues : elles permettent de répartir les données et les traitements sur plusieurs serveurs afin d'améliorer les performances des traitements et de rendre le système robuste à la panne d'un ou plusieurs des serveurs.

… ET NÉCESSAIRE AUTOMATISATION DES PROCESSUS STATISTIQUES

D'un point de vue statistique, le volume des données ne permet plus les traitements manuels qui pouvaient être effectués, en général par les enquêteurs, et qui doivent donc être automatisés. Trois exemples peuvent être donnés : être capable de classer un produit dans une nomenclature détaillée, identifier les relances commerciales et remplacer un produit lorsqu'il disparaît.

- Dans les données de caisses, **les produits sont identifiés par leur code-barres *(figure 1)*** et celui-ci ne donne pas d'information sur la nature du produit suivi. Compte tenu du nombre de codes-barres présents dans les données de caisses (près de 9 millions), il est impensable de rechercher manuellement à quel produit correspond chacun d'eux. Pour résoudre ce problème, l'Insee achète à un panéliste un dictionnaire de codes-barres, décrivant très précisément les caractéristiques du produit associé à chaque code-barres. **Classer les produits** revient alors à construire une simple table de passage entre ce dictionnaire et la nomenclature Coicop.

9. Plus précisément, seul 1,3 milliard est effectivement exploité pour l'IPC (du fait de l'exclusion de certains produits). Ces données sont ensuite consolidées par mois et par article de classes équivalentes.

- Ce dictionnaire de codes-barres permet également de traiter correctement le cas des « **relances commerciales** ». Ces relances consistent en une modification marginale du *packaging* du produit avec souvent un prix en hausse, ou un prix stable mais avec un volume de produit vendu plus faible. Ces relances commerciales masquent en général une hausse de prix, il est donc fondamental de pouvoir les repérer. Dans une collecte sur le terrain, c'est l'enquêteur qui identifie la relance ; avec les données de caisses, le code-barres change en général avec le *packaging* et il faut être capable, de manière automatique, de relier le produit initial à sa relance. En Suède, Tongur (2019) estime à 0,1 point le biais qui pourrait s'ensuivre si la relance n'était pas identifiée dans les données de caisses. Dans le cas français, l'existence d'un dictionnaire de codes-barres permet de faire le lien nécessaire entre un produit et sa relance commerciale et de bien enregistrer la hausse de prix associée.
- Le **remplacement des produits** appartenant au panier de l'IPC et disparaissant en cours d'année est également une opération impliquant fortement les enquêteurs. Ils choisissent le produit remplaçant de manière à être le plus proche possible du produit disparu et décident s'il est nécessaire ou non d'effectuer un ajustement qualité[10]. Cette opération stratégique pour l'IPC peut elle aussi être automatisée (Léonard *et alii,* 2017). Le produit remplaçant est tiré aléatoirement parmi les produits de la même variété et un ajustement qualité est réalisé systématiquement, en comparant le prix du produit remplacé et du produit remplaçant au cours d'une même période : la différence de qualité est estimée égale à la différence de prix. Cette méthode d'ajustement de la qualité est couramment utilisée dans l'IPC, mais les prix comparés sont presque toujours imputés car lorsqu'on recourt à des relevés de prix, il n'est en général pas possible de comparer au cours d'une même période le prix du produit disparu et remplaçant. Par définition, on n'a pas anticipé que le produit allait disparaître et on n'a pas relevé le prix du produit remplaçant avant même de savoir que le produit remplacé allait disparaître. Puisque les données de caisses sont exhaustives, on peut y rechercher *a posteriori* le prix passé d'un produit.

UN CHOIX UN PEU DIFFÉRENT DE NOS PARTENAIRES EUROPÉENS

L'achat d'un dictionnaire de codes-barres[11] et le recours à des technologies de *big data* permettent ainsi de conserver les concepts de l'IPC actuel (en particulier l'idée d'un panier fixe annuellement) tout en exploitant les données de caisses dans leur exhaustivité. Cette situation est assez singulière dans le paysage des instituts statistiques recourant aux données de caisses.

Historiquement, et c'est encore la solution retenue par un certain nombre de pays comme la Suède et l'Italie, les pays qui ont utilisé les données de caisses pour calculer l'IPC ont tiré un échantillon de produits de manière à pouvoir réaliser manuellement les trois traitements précédemment décrits : la classification, l'identification des relances, le remplacement des produits. Les concepts de l'IPC sont alors strictement conservés et on utilise les données de caisses comme base de sondage, afin de repérer le plus rapidement possible l'apparition ou la suppression de produits ou pour connaître précisément les poids associés à chaque produit.

10. Pour neutraliser le fait que le produit remplaçant peut être de qualité légèrement différente du produit disparu.
11. Les INS des autres pays européens n'ont pas un tel référentiel ; ils s'appuient en général sur une description (plutôt courte) du produit que l'on peut retrouver sur les tickets de caisses et à l'aide de méthodes de *machine learning* réussissent à classer les codes-barres dans la nomenclature de fonction utilisée pour le calcul de l'IPC.

Les développements suivants ont visé à bénéficier de toute la précision apportée par le caractère exhaustif des données de caisses, et à limiter les traitements manuels liés aux remplacements, qui deviennent très vite importants dès qu'on augmente la taille de l'échantillon. La méthode précédente a été ainsi progressivement remplacée par une exploitation des données de caisses qui renonçait à la fixité annuelle du panier de produits suivis : l'évolution mensuelle des prix a été mesurée sur un panel cylindré de produits présents au cours de deux mois consécutifs ; ces évolutions mensuelles étaient ensuite chaînées les unes aux autres. Ce chaînage d'indices à fréquence trop élevée est connu des statisticiens des prix pour créer des dérives d'indice. Pour éviter cela, il a fallu renoncer, au niveau le plus fin, aux pondérations[12], une innovation pourtant importante des données de caisses.

Une dernière génération d'indices enfin a émergé : les indices multilatéraux, méthode GEKS (Diewert, Fox Ivancic, 2009) ou Geary-Khamis (Chessa 2015), s'inspirent des méthodes utilisées pour la comparaison spatiale des prix et permettent de traiter le fait que les paniers puissent être différents chaque mois. Celles-ci sont néanmoins moins intuitives et plus difficiles à expliquer au grand public.

AU FINAL, QUEL IMPACT DES DONNÉES DE CAISSES ?

Les données de caisses seront utilisées pour la première fois pour calculer l'IPC publié en janvier 2020. 30 000 prix relevés mensuellement par les enquêteurs seront remplacés par environ 77 millions de produits présents dans les données de caisses. En dehors de ce champ, la collecte actuelle sera conservée.

> *« Les données de caisses seront utilisées pour la première fois pour calculer l'IPC publié en janvier 2020. »*

Avant d'utiliser cette nouvelle source de données dans la production d'une statistique aussi importante pour le débat public que l'IPC, il a été nécessaire de faire la preuve de sa fiabilité, de sa pérennité et de son apport à la mesure de l'inflation. Une expérimentation et des travaux méthodologiques ont permis de définir le traitement des données de caisses et leur intégration dans l'IPC (Leclair *et alii*, 2019). Une application informatique spécifique a été développée, en technologie *big data*, pour s'assurer de la réception des données, des contrôles statistiques et des traitements à réaliser : l'IPC est produit en effet dans des délais très restreints qui nécessitent de s'appuyer sur des traitements informatiques robustes. La mise à disposition des données de caisses a été fiabilisée par un arrêté (voir *supra*), doublé, pour certaines enseignes, de conventions.

Enfin, avant l'utilisation effective des données de caisses dans l'IPC, l'Insee a souhaité réaliser pendant une année une répétition générale : alors que l'IPC publié chaque mois en 2019 s'appuyait sur les relevés de prix des enquêteurs (et des autres sources traditionnelles utilisées par l'IPC), en parallèle, un IPC a été produit en utilisant les données de caisses.

12. Ce sont en effet les pondérations qui créent la dérive de l'indice : le poids dans la consommation dépendant en général du niveau de prix, une période de promotion lie un poids élevé à une baisse de prix alors que le retour au prix normal s'accompagne d'un poids dans la consommation faible. L'indice chaîné mensuellement et utilisant des pondérations ne revient pas ainsi à son niveau d'origine après une période de promotion, car il ne pondère pas de la même manière les hausses et les baisses de prix.

Cette répétition générale permet de roder le processus de production ; elle permet surtout de comparer les résultats. Quel impact alors de cette nouvelle source de données ? L'impact sur l'inflation d'ensemble est peu visible car au final, les données de caisses ne représentent qu'un poids faible de la consommation d'ensemble (11 % environ) : beaucoup de produits ne peuvent pas être suivis par des données de caisses (les services, les produits frais, en l'absence de codes-barres) ou ne le sont pas pour des raisons méthodologiques (l'habillement et les biens durables, pour lesquels la rotation des produits est importante et la méthodologie utilisée pour les ajustements qualité est spécifique) et enfin, la consommation des ménages se fait aussi dans d'autres formes de vente que les super et hypermarchés.

« L'impact sur l'inflation d'ensemble est peu visible car au final, les données de caisses ne représentent qu'un poids faible de la consommation d'ensemble (11 % environ). »

Mais à un niveau plus fin, pour des postes où les données de caisses sont plus fortement utilisées, des différences peuvent être notées. L'analyse fine des écarts montre qu'elles s'expliquent essentiellement par trois facteurs :

- une meilleure représentativité des produits suivis : la connaissance fine des quantités dans les données de caisses amène à suivre des variétés de produits qui ne l'étaient pas jusqu'à présent faute d'avoir pu identifier leur importance ; or, ces variétés de produits ont des dynamiques de prix propres qui n'étaient pas prises en compte préalablement ;
- une meilleure précision de l'indice : sur les mêmes variétés de produits, des différences d'évolution de prix peuvent exister du fait de l'imprécision de l'échantillonnage,
- un meilleur suivi des prix ; la meilleure intégration des promotions dans les données de caisses (voir *supra*) permet de mettre en avant des évolutions de prix qui ne peuvent l'être dans la collecte traditionnelle.

Les données de caisses sont donc une source prometteuse pour le calcul de statistiques de prix. En janvier 2020, seules les statistiques de prix publiées actuellement (l'IPC, l'indice des prix dans la grande distribution) seront produites. Mais dans un avenir plus lointain, de nouvelles statistiques pourront être diffusées : des prix moyens pour de nombreux produits, des comparaisons spatiales de prix ou des indices régionaux. Des études méthodologiques se poursuivront pour exploiter les données de caisses sur des champs pour lesquels elles ne sont pas encore mobilisées, l'habillement ou les biens durables par exemple. Des travaux seront menés pour accéder à de nouvelles données de caisses, celles des *hard-discounters* ou de la grande distribution spécialisée.

BIBLIOGRAPHIE

BLANCHET, Didier et GIVORD, Pauline, 2017. Données massives, statistique publique et mesure de l'économie. In : *L'Économie française, édition 2017.* [en ligne]. Insee Références, pp. 59-77. [Consulté le 7 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/2894010/ECOFRA17b_D1_big-data.pdf*

BOSKIN, Michael J., 1996. Toward a More Accurate Measure of the Cost of Living : Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index. In : *site de l'administration de la Sécurité Sociale des États-Unis.* [en ligne]. [Consulté le 7 octobre 2019]. Disponible à l'adresse : *http://www.ssa.gov/history/reports/boskinrpt.html*

CHESSA, Antonio, 2015. Towards a generic price index method for scanner data in the Dutch CPI. In : *Fourteenth Meeting of the Ottawa Group (International Working Group On Price Indices).* [en ligne]. 20-22 mai 2015. Tokyo, Japon. [Consulté le 7 octobre 2019]. Disponible à l'adresse : *https://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s1room2.pdf*

CLÉ, Émeline, JALUZOT, Laurence, MALAVAL, Fabien, RATEAU, Guillaume, SAUVADET, Luc, 2016. *En 2015, les prix en région parisienne dépassent de 9 % ceux de la province* [en ligne]. 14 avril 2016. Insee Première, n°1590. [Consulté le 7 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/version-html/1908158/ip1590.pdf*

DIEWERT, Erwin, FOX, Kevin J., IVANCIC, Lorraine, 2009. Scanner Data, Time Aggregation and the Construction of Price Indexes. In : *Eleventh Meeting of the Ottawa Group (International Working Group On Price Indices).* [en ligne]. 27-29 mai 2009. Neuchâtel, Suisse. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/a49bc2a164b232c4ca2576a100773522?OpenDocument*

EUROSTAT, 2017. *Practical Guide for Processing Supermarket Scanner Data*. [en ligne]. Septembre 2017. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf*

FMI, 2004. *Manuel des prix à la consommation. Théorie et pratique.* OIT/FMI/OCDE/CEE–ONU/Eurostat/Banque mondiale. Genève, Organisation internationale du travail. ISBN 1-58906-330-9

JALUZOT, Laurence et SILLARD, Patrick, 2016. *Échantillonnage des agglomérations de l'IPC pour la base 2015.* [en ligne]. Janvier 2016. Insee, Document de travail, N°F1601. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/2022137/F1601.pdf*

JANY-CATRICE, Florence, 2019. *L'indice des prix à la consommation.* Édition La Découverte. Collection Repères, N°717. Janvier 2019. ISBN 978-2-7071-9931-7

LECLAIR, Marie, LÉONARD, Isabelle, RATEAU, Guillaume, SILLARD, Patrick, VARLET, Gaëtan et VERNÉDAL, Pierre, 2019. Les données de caisses : avancées méthodologiques et nouveaux enjeux pour le calcul d'un indice des prix à la consommation. In : *Économie et statistique.* [en ligne] 17 septembre 2019. N°509, pp. 13-31. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4203515/509_Leclair-Leonard-Rateau-Sillard-Varlet-Vernedal-FR.pdf*

LECLAIR, Marie, et PASSERON, Vladimir, 2017. *Une inflation modérée depuis le passage à l'euro.* [en ligne]. 24 mai 2017. Insee Focus, N°87. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/2854085*

LÉONARD, Isabelle, SILLARD, Patrick, VARLET, Gaëtan, ZOYEM, Jean-Paul, 2017. *Scanner data and quality adjustment.* [en ligne]. Juin 2017. Insee, Document de travail, N°F1704. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/2912650/F1704.pdf*

LÉONARD, Isabelle, SILLARD, Patrick, VARLET, Gaëtan, ZOYEM, Jean-Paul, 2019. Écarts spatiaux de niveaux de prix entre régions et villes françaises avec des données de caisses. In : *Économie et statistique.* [en ligne]. 17 septembre 2019. N°509, pp. 73-87. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4203527/509_Leonard-Sillard-Varlet-Zoyem-FR.pdf*

LEQUILLER, François, 1997. L'indice des prix à la consommation surestime-t-il l'inflation ? In : *Économie et statistique.* [en ligne]. Mars 1997. N° 303, pp. 3-32. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.epsilon.insee.fr/jspui/bitstream/1/21417/1/estat_1997_303_1.pdf*

MOATI, Philippe et ROCHEFORT, Robert, 2008. *Mesurer le pouvoir d'achat.* [en ligne]. Janvier 2008. Édition La Documentation française, Collection Les Rapports du Conseil d'analyse économique. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.vie-publique.fr/sites/default/files/rapport/pdf/084000050.pdf*

RIVIÈRE, Pascal, 2018. Utiliser les déclarations administratives à des fins statistiques. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. N°N1, pp. 14-24. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647013/courstat-1-5.pdf*

SILLARD, Patrick, 2017. *Indices de prix à la consommation.* [en ligne]. 7 août 2017. Insee, Document de travail, N°F1706. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/version-html/2964204/F1706.pdf*

TASSI, Philippe, 2018. Les apports des Big Data. In : *Économie et statistique.* [en ligne]. N°505-506, pp. 5-15. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3705956/505-506_Tassi-FR.pdf*

TONGUR, Can, 2019. Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket. In : *Économie et statistique.* [en ligne]. 17 septembre 2019. N°509, pp. 33-50. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4203519/509_Tongur-FR.pdf*

UNECE, 2018. *Report of the Group of Experts on Consumer Price Indices.* [en ligne]. 7-9 mai 2018. Fourteenth session, Genève, Suisse. ECE/CES/GE.22/2018/2. [Consulté le 8 octobre 2019]. Disponible à l'adresse : *https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Report_of_CPI_expert_group_meeting_7-9_May_2018.pdf*