Les données de caisse : avancées méthodologiques et nouveaux enjeux pour le calcul d'un indice des prix à la consommation

Scanner Data: Advances in Methodology and New Challenges for Computing Consumer Price Indices

Marie Leclair*, Isabelle Léonard*, Guillaume Rateau, Patrick Sillard***, Gaëtan Varlet** et Pierre Vernédal******

Résumé – Lorsque les consommateurs passent à la caisse des magasins, les codes-barres (appelés également GTIN, pour *Global Trade Item Number*) des produits achetés sont scannés et les quantités achetées et les prix associés à chaque code-barres sont ainsi enregistrés. Ces données de caisse sont très prometteuses pour la construction des indices de prix à la consommation et pourraient se substituer ainsi à des relevés effectués par des enquêteurs. Le volume des données et les nouvelles informations qu'elles apportent nécessitent, à concepts inchangés de l'indice des prix à la consommation, de répondre à de nouvelles problématiques méthodologiques : l'agrégation des prix pour produire des indices, le traitement des ajustements qualité, le classement des produits par variété homogène de produits, le traitement des relances et des promotions, etc. L'article présente les orientations prises par la France face à ces nouvelles problématiques.

Abstract – *When consumers pay for their purchases at the store checkout, the barcodes (also known as GTINs, for Global Trade Item Number) of the goods purchased are scanned, recording quantities and the prices linked to each barcode in the process. Scanner data present an opportunity for use in constructing consumer price indices, which could supersede the use of survey data. Based on the existing concept of consumer price indices, the volume and new types of information provided by scanner datasets raise a number of new methodological questions, in particular in relation to price aggregation to produce indices, handling quality adjustments, classifying goods by homogeneous consumption segment and product relaunches and promotions. This article looks at how these questions have been addressed in France.*

Codes JEL / JEL Classification : E31, C8, D1
Mots-clés : indices des prix à la consommation, données de caisse
Keywords: consumer price indices, scanner data

** Insee, division des prix à la consommation (marie.leclair@insee.fr ; isabelle.leonard@insee.fr)*
*** Insee, division des prix à la consommation au moment de la rédaction de l'article (guillaume.rateau@enseignementsup.gouv.fr ; gaetan.varlet@insee.fr)*
**** Insee, département des méthodes statistiques (patrick.sillard@insee.fr)*
***** Insee, centre national d'informatique d'Orléans (pierre.vernedal@insee.fr)*

Nous remercions deux rapporteurs anonymes pour leurs commentaires et suggestions ainsi que Pascal Chevalier pour sa relecture.

Reçu le 16 octobre 2017, accepté après révisions le 26 juin 2018

Pour citer cet article : Leclair, M., Léonard, I., Rateau, G., Sillard, P., Varlet, G. & Vernédal, P. (2019). Scanner Data: Advances in Methodology and New Challenges for Computing Consumer Price Indices. *Economie et Statistique / Economics and Statistics*, 509, 13–29. https://doi.org/10.24187/ecostat.2019.509.1981

Lorsque les consommateurs paient leurs achats à la caisse des magasins, les codes-barres (appelés également GTIN, pour *Global Trade Item Number* ou EAN, pour *European Article Numbering*) des produits achetés sont scannés. Ce passage en caisse donne lieu à l'enregistrement des quantités achetées et des prix associés à chaque code-barres. Ces données que l'on appelle données de caisse, très volumineuses avec 1.7 milliards d'enregistrements par mois pour la grande distribution, sont centralisées et utilisées par les enseignes depuis de nombreuses années à des fins de gestion et d'étude de marché. Elles sont d'une richesse sans précédent pour le calcul des indices des prix à la consommation (IPC) : l'accès à ces données permet aux statisticiens de disposer de l'ensemble des prix, mais également des données de ventes, dans les super et hypermarchés, ce qui n'est évidemment pas le cas avec les méthodes de collecte classiques utilisées jusqu'à présent où des enquêteurs vont relever les prix dans des points de vente physique. Cette richesse d'information permet de construire un IPC plus précis, plus pertinent avec un niveau de détail plus important. Elle soulève aussi de nouveaux enjeux, notamment du fait du volume d'informations à traiter qui limite les interventions manuelles.

Le projet français d'utilisation des données de caisse pour le calcul de l'IPC vise à exploiter l'intégralité des informations mises à disposition des données de caisse tout en conservant la méthodologie et les concepts actuels de l'IPC. Ce faisant, les données de caisse constituent, du point de vue de l'IPC, une nouvelle source d'information dont l'usage ne devrait pas engendrer de rupture de série dans la mesure de l'inflation puisque les concepts de base demeurent. Ce choix qui n'est pas forcément celui des autres pays européens (qui ont oscillé à l'origine entre un échantillonnage des données de caisse pour reproduire les IPC actuels ou une modification des méthodes statistiques pour traiter l'important volume des données) nécessite toutefois de répondre, même à méthodologie constante, à de nouvelles questions statistiques.

Des questions centrales pour la construction des indices, comme le choix des formules d'agrégation permettant de passer de prix observés à un indice ou la manière de prendre en compte les changements de qualité des produits consommés, doivent trouver des réponses appropriées avec les données de caisse. L'article présente les différents choix effectués par le projet français d'utilisation des données de caisse à partir de janvier 2020, au regard de la définition actuelle de l'IPC. Les données de caisse exploitables à ce stade à des fins statistiques ne couvrent qu'une partie de la consommation des ménages[1], les produits d'alimentation industrielle, d'hygiène-beauté et d'entretien de la maison vendus en super et hypermarché. Pour le reste de la consommation (autres formes de vente, autres biens et services), la méthodologie utilisée jusqu'à présent dans l'IPC et les modalités de recueil actuel d'information sont conservées.

Des avancées méthodologiques permises par les données de caisse

Une amélioration de l'échantillonnage des produits suivis

L'IPC est un indice de type Laspeyres, à panier fixe chaîné annuellement : au cours d'une année, le principe de mesure consiste à suivre les prix de produits précis observés tous les mois dans les mêmes points de vente (encadré 1). On s'assure ainsi que l'évolution des prix mesurée n'est pas liée à un changement de la qualité des produits consommés. Les produits suivis doivent être représentatifs de la consommation des ménages. Si l'on disposait d'une connaissance exhaustive des transactions réalisées par les ménages, il serait possible de sélectionner les produits à inclure dans le panier de l'IPC par échantillonnage aléatoire. Dans l'approche classique, avant l'utilisation des données de caisse, en l'absence de cette information, on s'appuie sur une connaissance approximative des dépenses de consommation des ménages selon une nomenclature comprenant environ 300 regroupements élémentaires, appelés postes. Les poids relatifs des dépenses associées à chacun des postes sont fondés sur des informations rassemblées dans le cadre de la Comptabilité nationale. Dans ces conditions, l'échantillon est construit par une méthode de quotas : l'enquêteur de l'Insee choisit des produits dont il relèvera ensuite le prix chaque mois, en respectant un nombre de relevés par forme de vente et pour une variété donnée de produits. Ces quotas dépendent d'informations diverses (Comptabilité nationale pour le poids des différents postes de la consommation, sources professionnelles pour

1. En effet, si des données de caisse existent pour d'autres produits, elles ne sont pas à ce jour exploitables pour l'IPC car elles posent notamment des problèmes spécifiques de récupération de données (pas de centralisation unique des données), d'identification (pas de référentiel des codes-barres) et de remplacement (fort turnover des produits électroniques ou de l'habillement par exemple), voir encadré 3.

les formes de ventes ou les gammes de produits, etc.). Les unités urbaines dans lesquelles l'enquêteur relève ces prix sont quant à elles déterminées par un tirage aléatoire, à proportion de leur importance dans la consommation des ménages (Jaluzot & Sillard, 2016).

L'absence de base de sondage ne permet pas de procéder à un tirage aléatoire de l'échantillon et de réaliser un calcul de la précision de l'indice sans le recours à des hypothèses de sondage aléatoire. À l'inverse, les données de caisse (encadré 2) offrent une vue exhaustive des ventes par article précis, point de vente et jour de vente pour les hyper et supermarchés. En renonçant à un échantillonnage et en fondant le calcul de l'indice sur l'exhaustivité des ventes[2], le choix effectué ici, on tend à supprimer complètement cet aléa.

Une nouvelle manière d'agréger les prix pour construire des indices

En exploitant l'exhaustivité de l'information des données de caisse, de nouvelles questions se posent notamment sur l'agrégation des prix. Pour passer des prix élémentaires par produit à un indice d'ensemble, il faut en effet faire le choix de formules d'agrégation dont les conséquences sur l'indice produit ne sont pas anodines.

2. Plus précisément, les produits sélectionnés dans le panier données de caisse correspondent à l'ensemble des produits, classés dans une variété de produits, encore disponibles en décembre de l'année A–1 ; l'intégration des produits saisonniers, hors saison en décembre, doit encore être explorée. Les produits trop particuliers, qu'il est difficile de classer dans une variété de produit, et dont le suivi serait compliqué par la non-pérennité de la variété sont exclus du panier.

Encadré 1 – **L'indice des prix à la consommation**

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 représentatif de la consommation des ménages, les poids et le panier de produits suivis sont renouvelés chaque année : l'IPC est un indice chaîné annuellement. 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 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.

Il existe une version harmonisée de l'indice des prix à la consommation (IPCH), comparable avec les indices des prix des autres pays européens. Sa méthodologie, son champ, sa fréquence sont définis très finement par un règlement européen. C'est globalement la même méthodologie que l'IPC, à l'exception du concept de prix suivi (prix brut pour l'IPC, prix net, après remboursement de la sécurité sociale, pour l'IPCH) et du champ (hors produits non marchands pour l'IPC).

L'IPC est actuellement construit en se fondant sur deux types de sources : des relevés de prix effectués par des enquêteurs de l'Insee sur le terrain (de l'ordre de 200 000 relevés chaque mois dans des unités urbaines représentatives du territoire français) dans diverses formes de ventes (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, etc.), soit que des bases de données puissent être mobilisées pour calculer les évolutions de prix (données de la Cnam pour les services de santé, par exemple). L'IPC est représentatif de l'ensemble des biens et services monétaires marchands consommés par les ménages sur le territoire français. Cette consommation peut être déclinée selon une nomenclature internationale par fonction de consommation appelée COICOP (*Classification of Individual Consumption by Purpose*).

Les données de caisse ne sont pas mobilisables pour l'ensemble de la consommation des ménages : les services par exemple ne sont pas suivis par des codes-barres ; les produits frais n'ont pas de GTIN mais des codes-barres spécifiques à chaque point de ventes. Par ailleurs, toutes les formes de ventes ne collectent pas de manière centralisée l'information provenant de leurs caisses (les petites supérettes indépendantes par exemple) ou n'utilisent pas les codes-barres (les marchés). Enfin, certains produits sont plus complexes à suivre de manière automatisée (habillement, produits électroniques, etc.) du fait notamment des problématiques de remplacement de ces produits. De ce fait, le projet dans une première étape vise uniquement à prendre en compte pour la production de l'IPC les données de caisse des hyper et supermarchés, de France métropolitaine, pour les produits de l'alimentation industrielle (fonctions COICOP 01 et 021), de l'hygiène beauté et de l'entretien (fonctions 0561, 09342, 12132). En dehors de ce champ, la collecte IPC actuelle sera conservée.

ENCADRÉ 2 – **Les données de caisse**

Les bases de données de caisse existent depuis de nombreuses années dans les systèmes d'information des enseignes qui les utilisent pour la gestion des stocks et leur politique marketing. L'Insee les reçoit actuellement sous forme de données quotidiennes agrégées par point de vente et article. Sont fournis la quantité vendue d'un article dans un magasin (indépendamment du nombre de clients à l'origine des ventes), le chiffre d'affaires ainsi généré, un court descriptif de l'article et le classement de l'article dans la nomenclature propre à l'enseigne. Quand ils ne sont pas fournis, les prix sont obtenus par division du chiffre d'affaires par la quantité d'articles vendus.

Les points de vente sont repérés par un identifiant propre à l'enseigne et les articles par leur GTIN (*Global Trade Item Number*) ou par un identifiant propre à l'enseigne, voire au point de vente, matérialisé sur les articles par un code-barres. Le GTIN est un identifiant des articles manufacturés géré au niveau international par l'organisme GS1 dont le rôle est de faciliter la collaboration entre partenaires commerciaux, organisations et prestataires de technologies. À chaque article manufacturé correspond un GTIN et un seul sur une période de temps donnée. Pour compléter ces données de caisse, l'Insee achète à une société d'études de marché des référentiels d'articles et de points de vente. Les articles du référentiel sont très précisément décrits à l'aide d'une vingtaine de variables. Certaines variables sont communes à l'ensemble des familles (par exemple la marque du produit ou son volume) ; d'autres sont propres à chaque famille (par exemple le taux de matière grasse pour les yaourts). Ce référentiel couvre les produits de grande consommation dans les grandes surfaces alimentaires.

Les premières études méthodologiques ont été réalisées en 2011 sur les données, agrégées hebdomadairement, de dix-sept familles d'articles (yaourts, huiles, café, etc.) vendus dans un échantillon de 1 000 points de vente de métropole – hors Corse – appartenant à six enseignes. Ces données portaient sur les années 2007 à 2009. 45 à 50 millions d'observations ont été étudiées pour chacune des trois années. En raison de l'agrégation hebdomadaire, le prix étudié était issu d'une moyenne arithmétique des prix quotidiens pondérés par les quantités vendues. Les études sur les effets qualité ont été menées à partir de ces données.

À partir de 2013, les études ont été fondées sur les données quotidiennes de cinq enseignes représentant environ 30 % de part de marché.

À l'heure actuelle, le prix d'un produit donné n'est relevé qu'une fois par mois par un enquêteur de l'Insee. Pour éviter d'éventuels effets de grappe, c'est-à-dire une corrélation au sein d'un même point de vente des évolutions de prix, pour une variété donnée de produits, un seul prix est relevé au sein d'un même point de vente. Pour donner un exemple, au sein d'un supermarché A, le prix de la boîte de petit pois de 150 g de marque X n'est relevé que le premier jeudi du mois et aucune autre boîte de petit pois n'est relevée au cours du mois dans ce supermarché A. Par ailleurs, l'impossibilité de connaître le chiffre d'affaires associé à chacun des produits conduit à équipondérer les articles d'une même variété suivis dans une agglomération donnée.

Les données de caisse offrent une information infiniment plus précise concernant les transactions ; plus de prix sont collectés et on dispose d'une information sur le poids, dans les dépenses, de chaque produit : les chiffres d'affaires et les quantités vendues dans les hyper et supermarchés et donc les prix moyens pratiqués chaque jour sont en effet connus dans chaque magasin et pour chaque article (les prix de toutes les boîtes de petits pois sont connus pour tous les jours où il y a des ventes). Il est donc envisageable d'adapter les formules d'agrégation des prix relevés pour se rapprocher des concepts idéaux : agrégation des prix d'une variété de produits donnée entre points de vente (agrégation spatiale, le prix des boîtes de petits pois vendues dans différents magasins), mais aussi au sein du point de vente (agrégation des produits, l'ensemble des boîtes de petits pois, quelle que soit la marque, vendues dans un magasin donné) et également pour un produit donné, agrégation temporelle puisque le prix est connu à différents moments du mois (les différents prix de la boîte de petits pois de la marque X sont observés à différents moments du mois). Les deux derniers types d'agrégation ne sont pas praticables dans le cadre de la collecte classique de l'IPC à partir de données collectées par les enquêteurs.

Agrégation de la dimension spatiale et des produits

Actuellement, puisque pour une variété de produits donnée, un seul relevé de prix est effectué au cours du mois et dans un point de vente donné, la première cellule d'agrégation consiste à agréger des prix relevés dans différents points de ventes pour une même variété de produit et une même agglomération. En l'absence d'information fine sur la consommation (le poids des ventes de petits pois dans le supermarché A

par rapport à celles effectuées dans le supermarché B), ces prix sont équipondérés. À ce niveau, deux formules d'agrégation des prix sont retenues par les standards internationaux (FMI, 2004 ; Eurostat, 2013) et sont toutes deux utilisées pour construire l'IPC français :

1) l'indice de Dutot ($I^D_{k,m}$), avec lequel l'évolution des prix est mesurée par le rapport de prix moyens entre différents mois de l'année, ces prix moyens étant calculés par une moyenne arithmétique simple des prix collectés dans chaque unité urbaine ; $I^D_{k,m} = \frac{\sum_{i \in K} p_{i,m}}{\sum_{i \in K} p_{i,0}}$ où $p_{i,m}$ est le prix du produit *i* appartenant à la variété *k* au cours du mois *m* ;

2) l'indice de Jevons ($I^J_{k,m}$), c'est-à-dire une moyenne géométrique des évolutions de prix entre deux mois $I^J_{k,m} = \prod_{i \in K} \left(\frac{p_{i,m}}{p_{i,0}} \right)^{1/n}$, avec *n* le nombre d'observations de produits pour la variété *k*.

Le choix de l'une ou l'autre des formules tient à la fois à des critères statistiques et à des considérations économiques. L'indice de Dutot, plus intuitif pour le grand public, tend à donner plus de poids aux produits dont les prix sont élevés et n'est donc pas très pertinent pour rendre compte de l'évolution moyenne des prix de produits hétérogènes, regroupant des produits de qualité diverse, comme les lave-linge par exemple, pour lesquels la dispersion des niveaux de prix est importante. À l'inverse, l'indice de Jevons est mieux adapté car il gomme les effets de dispersion. Lorsque les variétés de produits sont homogènes, avec peu de variations de caractéristiques ou de qualité d'un produit à l'autre, (comme la baguette de pain), alors l'usage de l'indice de Dutot, plus intuitif, se justifie.

La théorie économique est également un recours pour déterminer quelle formule est adaptée (Sillard, 2017) : un indice de Dutot est cohérent avec une fonction d'utilité du consommateur de type Leontief (sans substitution entre les produits consommés) tandis que les indices de Jevons correspondent à des fonctions de type Cobb-Douglas[3] (avec élasticité de substitution unitaire entre les produits). Dans la configuration actuelle du calcul de l'IPC, un seul prix pour une variété de produits donnée est relevé dans un point de vente particulier. Avec les formules de Dutot pour les variétés homogènes et de Jevons pour les variétés hétérogènes, on fait l'hypothèse implicite qu'il n'y a pas de substitution entre points de vente pour des produits homogènes tandis qu'il y en a pour les produits hétérogènes. En d'autres termes, le consommateur effectue ses arbitrages de prix à l'échelle de l'agglomération pour les variétés hétérogènes de produits (les lave-linge) et à l'échelle des points de vente pour les produits homogènes (la baguette de pain). À un niveau plus agrégé, où les poids sont connus (poids des agglomérations dans la consommation des ménages, poids de la variété dans la consommation des ménages), l'agrégation est de type Laspeyres arithmétique.

Avec les données de caisse, le choix de ces indices élémentaires est modifié : il y a d'une part plus de prix observés impliquant potentiellement plus de substitution (plus d'un produit d'une variété donnée au sein d'un point de vente) et d'autre part, les poids des ventes de chaque produit et de chaque point de vente sont connus, permettant de s'abstraire de l'équipondération des formules de Dutot et de Jevons.

Différentes formules d'indice ont donc été considérées : elles consistent à retenir des indices de type Laspeyres, arithmétiques ou géométriques, selon le niveau d'agrégation (entre produits d'une même variété au sein du point de vente, entre points de vente pour une même variété, entre variétés), utilisant comme pondération le poids dans les ventes telles qu'observées dans les données de caisse[4]. Le choix entre un Laspeyres arithmétique ou géométrique n'est pas anodin pour la mesure de l'inflation. En termes de comportement micro-économique du consommateur, la moyenne géométrique repose sur une hypothèse de substituabilité des produits tandis que la moyenne arithmétique suppose que les produits sont complémentaires. Si le prix d'un bien diminue relativement à celui des autres biens, lorsque les biens sont substituables, le consommateur va acheter davantage du bien dont le prix a diminué et réduire sa consommation des autres biens. De ce fait, plus les produits sont substituables, plus le consommateur bénéficie de la baisse des prix. Si, à l'inverse, les produits ne sont pas substituables, il ne bénéficie de la baisse

3. L'indice s'écrit comme le rapport des coûts optimaux des paniers associés aux deux mois comparés. Le programme d'optimisation du consommateur est écrit à utilité constante dont le niveau est arbitraire puisque l'expression de l'indice en est indépendante. L'indice de Dutot peut en effet être obtenu, de la même façon, en considérant une utilité de Leontief.
4. Le poids est considéré sur l'ensemble de l'année A–1 tandis que le prix de base est celui de décembre.

de prix qu'en proportion de sa consommation (constante) du bien dont le prix baisse. Le choix des formules a donc des conséquences sur l'indice puisque l'impact de la baisse du prix d'un produit est plus important avec un indice géométrique qu'avec un indice arithmétique.

Le choix de la formule a été fait en fonction du comportement supposé du consommateur mais aussi de manière à exploiter l'information nouvelle apportée par les données de caisse sans modifier pour autant les hypothèses sous-jacentes à la construction de l'indice actuel. La possibilité pour le consommateur de substituer entre produits dépend (i) d'une part du fait que ces produits lui permettent de satisfaire les mêmes besoins et (ii) d'autre part de sa connaissance des différents prix pratiqués pour les différents produits et dans les différents points de vente.

Sur le premier point (i), la définition des variétés de produits permettant de satisfaire un même besoin se fait par expertise et on verra plus loin que les données de caisse et l'extension de la couverture des produits qu'elle implique sont à la fois un apport et une difficulté pour la définition de ces variétés du fait de la volumétrie des données à traiter (encadré 3, pour les difficultés informatiques à traiter un si gros volume de données). Ces variétés de consommation sont définies de manière à vérifier l'hypothèse qu'il n'y a pas de substitution entre variétés. Au-delà de cette agrégation élémentaire par variété, l'agrégation entre variétés de produits est de type Laspeyres pondéré.

Sur le second point (ii), obtenir une information sur les différents prix pratiqués afin d'arbitrer et de substituer entre produits se traduit rapidement pour le consommateur par un coût de prospection et de transport non négligeable. Différentes hypothèses peuvent être faites : on peut considérer que le consommateur dispose de cette information à coût quasi-nul au sein d'un point de vente (1), dans une unité urbaine (2) ou même, hypothèse extrême, pour l'ensemble de la France métropolitaine (3). En cohérence avec ces hypothèses alternatives, des indices de prix de yaourts vendus en supermarchés entre décembre 2008 et décembre 2009 (tableau 1) ont été construits selon 4 formules : (1) un indice de Laspeyres géométrique au sein d'un point de vente et arithmétique pour les niveaux supérieurs, (2) un indice de Laspeyres géométrique au sein d'une unité urbaine et arithmétique pour les niveaux supérieurs, (3) un indice de Laspeyres géométrique pour l'ensemble

Encadré 3 – **Un choix technique pour garantir de manière pérenne le traitement des gros volumes des données de caisse**

Les études présentées dans l'article ont été réalisées à l'aide de technologies informatiques « classiques ». En conséquence, compte-tenu des temps de traitement, elles sont en général appliquées à quelques variétés emblématiques de produits. La production mensuelle d'un IPC requiert de traiter l'intégralité du champ, soit un volume très important de données dans des délais très courts (une première estimation de l'IPC du mois m est publiée le dernier jour ouvré du même mois). À l'issue de tests, les bases de données classiques (relationnelles) n'ont pas été jugées capables de satisfaire de telles contraintes.

Les technologies qui ont émergé avec le phénomène des Big Data, en particulier le système Hadoop, permettent la maîtrise des temps de traitements de gros volumes de données. La nouveauté, en regard des bases de données relationnelles, réside dans la répartition des données et des traitements sur plusieurs serveurs. Ceci implique de pouvoir décomposer un traitement, par exemple une requête écrite en SQL, en un traitement exécuté sur chaque morceau de données (appelé « map ») et un traitement (appelé « reduce ») effectuant la synthèse des résultats « map ». Le moteur Hadoop qui s'en charge est écrit en java. Pour rendre ceci possible, les contraintes d'intégrités (clé primaire, clés étrangères), utilisées dans les bases de données relationnelles pour garantir des cohérences entre données, ont été abolies dans les systèmes Big Data, qui concernent davantage des entrepôts où les données s'empilent et sont moins sujettes à des modifications ponctuelles.

La délégation des traitements permet de contrôler les performances via l'augmentation du nombre de serveurs délégués (appelés « datanodes »). Les performances dépendent de manière linéaire des volumes à traiter et sont fonction du nombre de datanodes utilisés. Le système est robuste, une panne d'un datanode n'interrompant pas un traitement : Hadoop duplique chaque nouveau paquet de données sur au moins 3 datanodes ; ainsi lorsqu'un datanode est défaillant, Hadoop va réaffecter la tâche qui lui incombait à un datanode possédant un réplicat ce qui permet au traitement global d'aboutir normalement.

Hadoop est donc privilégié pour les développements « données de caisse » brassant les gros volumes, les données « synthétiques » résultantes sont ensuite injectées dans une base de données relationnelle où consultation de tableaux de bords et travaux de gestion s'effectuent dans le cadre d'une application « classique ».

de la France métropolitaine, (4) un indice de Laspeyres arithmétique au sein des points de ventes et pour tous les niveaux d'agrégation supérieurs. L'écart sur le glissement annuel du prix des yaourts est de 0.65 point de pourcentage selon les deux hypothèses extrêmes d'une substitution au sein de la France métropolitaine (3) et d'une absence de substitution, y compris au sein des points de vente (4).

Parmi ces configurations et pour des produits du type des yaourts, il paraît vraisemblable que, dans l'instantanéité de l'achat, le consommateur arbitre sur les prix, essentiellement parmi les produits vendus au sein du point de vente considéré et non entre différents points de vente. En effet, pour arbitrer entre points de vente, il faudrait que, dans un laps de temps court (celui de l'achat), le consommateur puisse se mettre en situation d'information complète sur les prix et parcoure les différents points de vente de son quartier pour procéder aux arbitrages requis. Pour des produits à faibles coûts de transaction (i.e. les variétés homogènes de produits), cette démarche est peu vraisemblable. Par conséquent, l'indice retenu *in fine* agrège les produits d'une même variété et dans un même point de vente à l'aide d'une formule de Laspeyres géométrique et les niveaux supérieurs à l'aide d'un Laspeyres arithmétique. Le choix de cette configuration est d'ailleurs cohérent avec l'agrégation pratiquée dans l'IPC actuellement. En effet, à l'heure actuelle, si le problème de l'agrégation au sein d'un point de vente ne se pose pas puisqu'un seul prix est relevé chaque mois dans un point de vente pour une variété donnée, la plupart des produits couverts par les données de caisse appartiennent à des variétés homogènes et font l'objet d'un calcul d'indice de Dutot à l'échelle de l'agglomération.

Agrégation temporelle

Dans le cadre actuel de l'IPC, les prix d'un produit ne sont relevés qu'une fois par mois pour un point de vente donné et une variété de produits donnée. La répartition de la collecte sur l'ensemble du mois permet, par échantillonnage et agrégation, de traiter l'évolution mensuelle des prix sans être dépendant d'une journée particulière du mois. Avec les données de caisse, on dispose de données de transactions détaillées par jour. Ce détail temporel des prix au cours du mois représente un surplus de données qu'il convient d'agréger pour obtenir un indice mensuel.

L'agrégation temporelle est un peu différente de l'agrégation des produits. Il est reconnu que pour agréger les prix de produits quasiment identiques, il est préférable de considérer les valeurs unitaires, autrement dit de prendre chaque mois la moyenne des prix pondérée par les quantités vendues (FMI, 2004). Toutefois, lorsque les produits diffèrent en nature ou en qualité, la méthode conduit à des biais importants. Dans la pratique actuelle de l'IPC, les quantités vendues sont inconnues à ce niveau de détail, si bien que cette méthode est inenvisageable. Les données de caisse, en revanche, donnent accès à cette information, et leur construction (chiffres d'affaires et quantités vendues) amène naturellement à effectuer ce calcul. La plupart des pays européens disposent de données mensuelles ou au mieux hebdomadaires, de sorte que cette méthode s'impose quasiment par nécessité[5] (encadré 4). Au cours

5. Cette méthode a également l'avantage de traiter implicitement des prix manquants. En effet, si un produit n'est pas vendu un jour donné, aucune information n'est disponible ce jour-là dans les données de caisse. Le suivi quotidien des prix implique donc de les imputer. Avec la valeur unitaire, l'imputation est implicite puisque ce jour-là on pondère par zéro le prix non observé.

Tableau 1
Glissement annuel de l'indice des prix des yaourts selon différentes formules d'agrégation, en 2009

Champ de substitution	Nombre de micro-indices	Glissement annuel (en %) (écart-type)
Variété (3)	9	- 4.29 % (0.16)
Variété × unité urbaine (2)	1 280	- 4.06 % (0.15)
Variété × point de vente (1)	2 335	- 3.87 % (0.15)
Aucun (4)	3 592	- 3.64 % (0.15)

Note : écart-type estimé par boostrap (100 réplications) ; le nombre de micro-indices correspond au nombre d'indices mesurés à l'aide d'un Laspeyres géométrique, qui font ensuite l'objet d'une agrégation de type Laspeyres arithmétique pour donner le glissement annuel. Lorsque le champ de substitution est la variété, les prix des yaourts sont agrégés par une moyenne géométrique en fonction des 9 variétés de yaourt définies. Ces 9 micro-indices sont ensuite agrégés selon une agrégation de type Laspeyres arithmétique. Selon ces formules d'agrégation, le prix des yaourts a baissé de 4.29 % entre décembre 2008 et décembre 2009. L'écart-type de cette évolution est estimé à 0.16 point.
Champ : échantillon de 3 592 yaourts répartis en 9 variétés de yaourts.
Source : données de caisse, 2008-2009.

ENCADRÉ 4 – **Les expériences européennes d'utilisation des données de caisse**

En Europe, quasiment tous les instituts statistiques ont actuellement lancé un projet visant à introduire l'utilisation des données de caisse dans la production de leur indice de prix. L'état d'avancement de ces projets est toutefois très contrasté. On dénombre neuf pays qui ont intégré le traitement de ces données dans leur système de production. L'institut des Pays-Bas (CBS) fait figure de précurseur et a débuté cette exploitation dès 2002, suivi par ceux de la Norvège en 2005, la Suisse (2008), la Suède (2012), la Belgique (2015), du Danemark (2016), de l'Islande (2016), du Luxembourg (2018) et de l'Italie (2018).

La plupart des pays reçoivent des données de transactions détaillées par code-barres et par point de vente, mais consolidées par semaine, limitant leur utilisation pour l'IPC à seulement deux ou trois semaines au cours du mois. Ces données sont accompagnées de différents systèmes de classification généralement propres à chaque distributeur. Les caractéristiques doivent quasi systématiquement être extraites du libellé inscrit sur les tickets de caisse décrivant le produit. En la matière, le projet de l'Insee fait figure d'exception avec l'accès à des données journalières, documentées de manière structurée suivant de nombreuses caractéristiques.

Sans un référentiel des codes-barres structuré, comme dans le cas français, la définition des variétés et leur classification au sein de la COICOP s'avèrent particulièrement difficiles. Elles reposent sur le système de classement plus ou moins complexe des distributeurs des articles ; l'extraction de l'information contenue dans le libellé des tickets s'appuie sur des techniques d'apprentissage et de *text mining*. Au niveau le plus fin, l'exploitation d'identifiants définis par les distributeurs, tels que les unités de gestion des stocks, permet de regrouper les codes-barres semblables et de rattacher les promotions fabricants aux articles originaux. La détection des relances est moins évidente et se fait indirectement en analysant les trajectoires des chiffres d'affaires et des quantités vendues et en essayant de détecter des substitutions.

Depuis leur début, les Pays-Bas ont mis en exploitation deux versions du traitement des données de caisse. Ces versions illustrent les différentes approches envisagées et leurs difficultés. Une de ces versions a notamment consisté à utiliser un panier annuel fixe et à agréger les prix au niveau de la variété par une moyenne géométrique. Alors que les indices produits étaient de qualité satisfaisante, le travail de maintien de l'échantillon et notamment le choix des produits remplaçants se sont révélés intenables en l'absence d'une description structurée des codes-barres.

Par la suite, les travaux méthodologiques se sont concentrés sur l'utilisation de paniers pouvant être renouvelés chaque mois. Ces paniers, dit dynamiques, permettent de s'affranchir du travail de remplacement au cas par cas. Seuls les produits les mieux vendus sont par ailleurs retenus dans le panier. Dans ces conditions, les indices élémentaires (permettant d'agréger les prix pour une même variété) sont des indices de Jevons chaînés tous les mois. Ce corpus méthodologique sert de base à la plupart des méthodes de traitement des données de caisse utilisées en Europe, notamment par les Pays-Bas, la Norvège, la Belgique et le Luxembourg. Il prend également une place importante dans les recommandations définies par Eurostat pour le traitement des données de caisse (Eurostat, 2017).

Dans cette méthode, les quantités vendues par produits à un niveau fin ne sont pas utilisées pour la construction de l'indice. Du fait du chaînage mensuel (le panier est renouvelé chaque mois), l'utilisation de ces quantités provoque une dérive de l'indice généralement spectaculaire. Pour éviter le phénomène de dérive du chaînage mensuel et exploiter l'information nouvelle des données de caisse sur les pondérations, de nouvelles méthodes sont considérées s'appuyant sur des méthodologies employées habituellement dans le cadre de la comparaison spatiale : méthode GEKS (Diewert *et al*., 2009), Geary-Khamis (Chessa, 2015). Ces méthodes permettent en effet de former un système transitif d'indices de prix. Avec de tels indices, l'introduction de l'information sur un nouveau mois modifie toutefois l'analyse que l'on fait du passé. Cette propriété n'est pas souhaitable pour construire des indices de prix qui dans de nombreux pays ne sont pas révisables. Pour s'abstraire de ces révisions, le principe est de travailler avec une fenêtre glissante comptant 13 ou 14 mois et de se contenter d'y appliquer le traitement de transitivité, sans pour autant rendre l'indice vraiment transitif sur tous les mois de l'année (Diewert & Fox, 2017 ; von der Lippe, 2012). Une autre approche pour asseoir une agrégation des prix des paniers dynamiques est plus axiomatique et cherche à déterminer la forme fonctionnelle optimale adaptée à ce contexte (Zhang *et al*., 2017).

du mois, cette agrégation est valide si le produit vendu est jugé identique quel que soit le jour de vente. Si ce n'est pas le cas, le bien doit être considéré comme un produit différent selon le jour où il est vendu. L'agrégation des prix des produits par jour s'apparente alors à l'agrégation de produits différents (cf. *supra*).

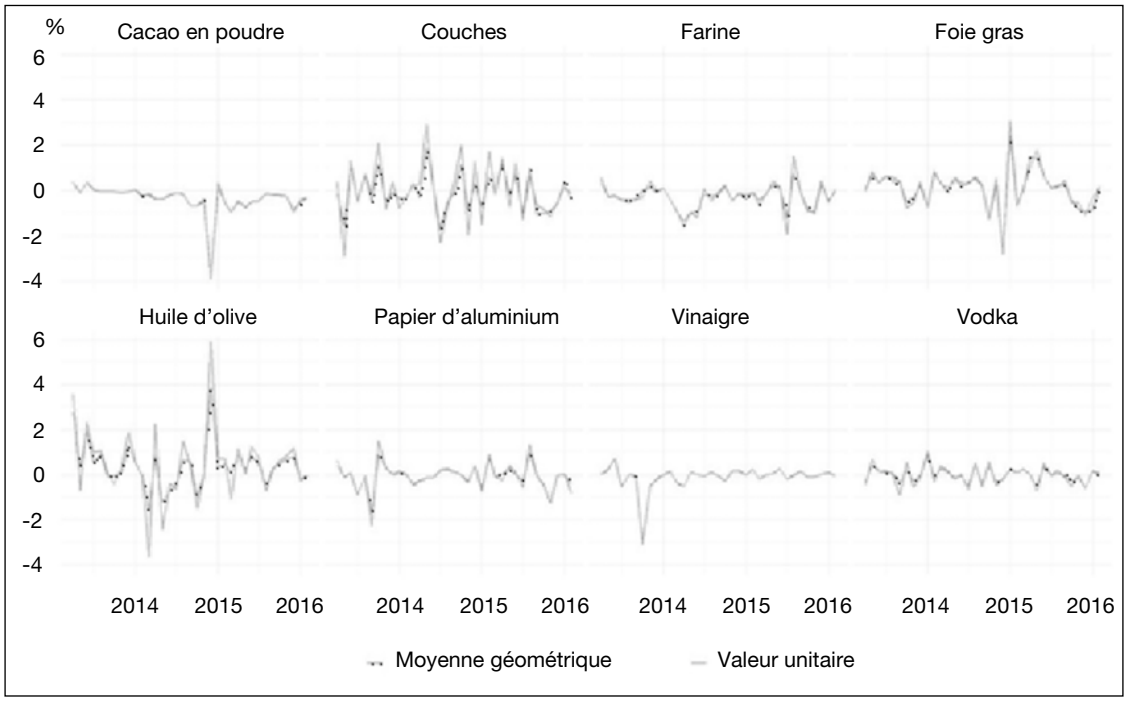
Le choix d'une formule plutôt qu'une autre a, là encore, un impact sensible sur les résultats obtenus en termes d'indice. Pour huit variétés de produits représentatives, des indices ont été construits entre 2013 et 2016 en agrégeant temporellement les prix soit grâce à une valeur unitaire ($\overline{p} = \sum_{i=1}^{28} v_{m,i} / \sum_{i=1}^{28} q_{m,i}$ avec v la dépense le jour j et q les quantités vendues le jour i), soit *via* une moyenne géométrique avec équipondération des jours au cours du mois ($\overline{p} = \prod_{i=1}^{28} p_{m,i}$ avec p le prix observé le jour i). Pour certaines

variétés de produits (les couches pour bébé, l'huile d'olive et, dans une moindre mesure, la farine de blé), les écarts entre les deux indices peuvent atteindre certains mois plusieurs points d'indice (figure I). L'utilisation des quantités courantes achetées dans le cadre de la formule de la valeur unitaire conduit à des indices plus volatiles. L'analyse fine de ces écarts sur l'huile d'olive montre qu'ils sont essentiellement provoqués par l'existence d'un petit volume de promotions magasins, de très courte durée et correspondant à un niveau de remise modéré. Durant ces promotions, les quantités vendues sont multipliées par un facteur allant fréquemment de 2 à 10. Dans un contexte de relative stabilité des prix, ces promotions participent activement à la dynamique à court terme des prix. Avec l'emploi de la formule de la valeur unitaire, l'impact de ces promotions sur les achats des ménages est mieux pris en compte, et la dynamique associée est plus visible dans les indices.

Pour choisir entre les deux formules, il convient de savoir si le jour de vente fait partie des caractéristiques du produit susceptibles d'en modifier l'utilité pour le consommateur. Pour certains produits suivis par l'IPC, notamment des services, le jour peut sembler une caractéristique importante du produit. Une nuit d'hôtel ou un billet de train un jour de week-end ou un jour de semaine sont des produits clairement distincts. Pour les produits suivis dans le champ de données de caisse, cette différence de produits en fonction du jour est bien moins évidente. On peut imaginer que le consommateur préfère faire ses courses certains jours de la semaine (week-end, mercredi ou lundi) et que, en réaction, les enseignes pourraient proposer systématiquement des promotions les jours les moins fréquentés. Ces différences de prix en fonction du jour ou même du moment de la journée sont observables, par exemple, dans le cas du commerce en ligne. Or, avec l'apparition de système d'affichage électronique des prix dans les magasins, les prix peuvent être modifiés rapidement et à faible coût.

L'existence d'une telle variation régulière de prix selon le jour de la semaine a été recherchée dans les données de caisse dont on dispose pour 2013 à 2015 pour huit variétés de produits (figure II). Sur cette période et pour les enseignes du champ, les résidus des moyennes mobiles de

Figure I
Glissement mensuel de l'indice des prix pour 8 variétés de produits selon deux formules d'agrégations temporelles, en %, de 2013 à 2016



Note : la valeur unitaire est le ratio des ventes du mois d'un produit et des quantités vendues au cours de même mois ; la moyenne géométrique pondère chaque prix quotidien du mois par le même poids.
Champ : prix des produits représentant les 8 variétés présentées.
Source : données de caisse de 4 enseignes représentant 30 % du marché, de 2013 à 2015.

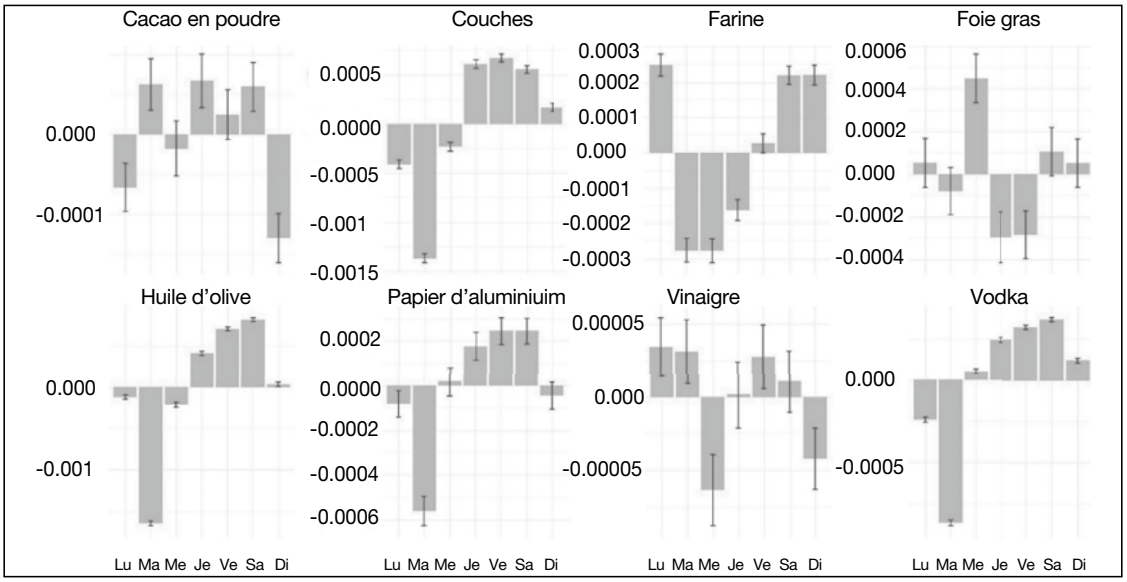
prix calculés sur une semaine montrent que les écarts de prix pour ces variétés entre jours de la semaine sont très faibles (les plus forts écarts sont de l'ordre de 0.1 %), et que pour ce type de produits, il n'y a pas eu de politique de fixation des prix différenciée au cours de la semaine sur la période considérée par les enseignes concernées.

Une amélioration du traitement des effets qualité

Pour la construction d'un IPC, le traitement des effets qualité est une question centrale, sujet à de nombreux débats. L'IPC est un indice à panier fixe chaîné annuellement. Au cours d'une année, les mêmes produits sont suivis chaque mois dans les mêmes points de vente. La constitution d'un panier fixe même annuel est bien sûr une gageure : des produits nouveaux apparaissent en cours d'année et des produits disparaissent. Pour assurer la continuité du panier tout au long de l'année et une mesure de l'évolution pure des prix (i.e. à « qualité » constante), les produits disparus sont remplacés par des produits proches et un ajustement qualité est effectué pour distinguer dans l'évolution de prix entre le produit disparu et le produit remplaçant, ce qui relève d'une pure évolution de prix de ce qui relève d'un changement de caractéristiques du produit. Différentes méthodes existent pour ajuster de la qualité : les méthodes de recouvrement et leurs différentes variantes (*bridged overlap*) sont les plus courantes et consistent à mesurer implicitement la différence de qualité par la différence de prix observée (conformément à la théorie économique dite « des références révélées ») ; l'option *pricing* qui repose sur une mesure à dire d'expert ; une modélisation, dite « hédonique », du prix en fonction des caractéristiques observables des produits (FMI, 2004, ch. 7). Parfois, lorsque le produit remplacé et le produit remplaçant sont jugés de qualité équivalente, aucun ajustement n'est pratiqué.

L'utilisation des données de caisse ne modifie pas sensiblement cette difficulté. D'une certaine manière, il l'atténue puisque la connaissance exhaustive des dépenses de consommation permet d'identifier plus rapidement la disparition d'un produit et de choisir un remplaçant dans le panier annuel ; elle rend également aisée la mesure simultanée des prix des produits remplaçants et remplacés puisque ceux-ci sont enregistrés dans les bases de données. La mécanique du choix du remplaçant doit toutefois être revisitée. Dans la pratique actuelle, un échantillon de produits seulement est suivi et la consigne donnée aux enquêteurs est de suivre des produits « bien vendus, bien

Figure II
Effet du jour de la semaine sur les prix observés de 2013 à 2015



Note de lecture : le dimanche, le prix du cacao en poudre est en moyenne inférieur de 0.01 % aux prix observé les autres jours.
Note : moyenne pondérée des résidus des moyennes mobiles calculées sur une semaine en gris, écart-type en trait plein noir.
Champ : prix des produits représentant les 8 variétés.
Source : données de caisse de 4 enseignes représentant 30 % du marché, de 2013 à 2015.

suivis » à la fois pour être le plus représentatif des produits consommés par les ménages et pour s'assurer que l'on pourra suivre les prix dans le temps, limitant ainsi les remplacements. Dans les données de caisse, le choix est de retenir l'exhaustivité des ventes : la rotation des produits et la taille du panier accroissent le nombre de disparitions et de remplacements au cours d'une année. Le volume de données à traiter ne permet pas de fonder le choix des produits remplaçants sur l'expertise humaine, comme actuellement. Un processus de décision automatisé est donc à construire.

Choix du produit remplaçant

À partir des données de 17 familles de produits, deux algorithmes de choix des produits remplaçants ont été testés : un algorithme déterministe et un algorithme alternatif, fondé en partie sur une sélection aléatoire.

Dans l'algorithme déterministe, le produit remplaçant est recherché dans la même variété de produits, le même point de vente et la même marque/gamme. En cas d'échec, si aucun produit vendu ne correspond à ces critères, le critère de la marque est relâché et la recherche s'effectue au sein de la variété et du point de vente. En cas de nouvel échec, la recherche est élargie à l'agglomération : même variété, même agglomération et même marque. Si besoin, le critère de la marque est à nouveau relâché, puis le critère géographique et enfin la recherche s'effectue au sein de la variété sur l'ensemble du territoire métropolitain. À une étape donnée, si plusieurs produits sont candidats, celui dont le prix, le mois précédent, est le plus proche du prix du produit à remplacer est retenu. Si des ex-æquo subsistent encore, le produit retenu est celui dont la quantité vendue est la plus proche de celle du produit à remplacer.

L'algorithme alternatif consiste simplement à rechercher le produit remplaçant parmi les articles de la même variété vendus dans le même magasin. Dans les très rares cas (moins de 0.1 %) où aucun produit n'est sélectionné, le critère de lieu est relâché par étape : même agglomération puis France métropolitaine si nécessaire (tableau 2). Cette recherche aboutit généralement à sélectionner un ensemble de produits « candidats » parmi lesquels le produit remplaçant est sélectionné aléatoirement. Cet algorithme est naturellement beaucoup plus simple à mettre en œuvre. Il est aussi plus fruste sur le plan économique. Les tests menés permettent d'étudier l'impact de ces différentes modalités de choix du remplaçant sur les indices de prix calculés (cf. *infra*).

Mesure de l'effet qualité

Une fois le produit remplaçant sélectionné, un ajustement qualité doit être effectué pour mesurer la différence de prix entre les deux produits, remplacé et remplaçant, due à la différence de caractéristiques des produits. Sont testées des méthodes usuelles adaptées au cas particulier des données de caisse. Par exemple, les méthodes par recouvrement reposent sur l'hypothèse qu'une différence de prix observée à un moment donné reflète une différence de qualité des produits. Dans la pratique actuelle de l'IPC, cette différence de prix « à un moment donné » doit être estimée car l'information sur le prix du produit remplacé et remplaçant porte sur deux dates différentes – on n'a en général aucune

Tableau 2
Type de remplacement, par famille de produits en 2009

(En %)

Type	Critères	Yaourts	Tablettes de chocolat	Fromage à pâte persillée	Œufs de poule	Café moulu à caféine
1	Même variété, point de vente, marque	73.0	55.7	58.0	16.9	33.8
2	Même variété, même point de vente	26.9	44.3	42.0	80.2	66.2
3	Même variété, agglomération, marque	0.0	0.0	0.0	2.8	0.0
4	Même variété, même agglomération	0.0	0.0	0.0	0.0	0.0
5	Même variété, même marque	0.0	0.0	0.0	0.0	0.0
6	Même variété	0.0	0.0	0.0	0.1	0.0
Ensemble		100.0	100.0	100.0	100.0	100.0

Note de lecture : 73 % des articles de type « yaourt » qui ont disparu au cours de l'année 2009 trouvent un remplaçant de même marque dans le même point de vente.
Source : échantillon de données de caisse pour 17 familles de produits dans 1 000 hyper et supermarchés.

information sur le prix du produit remplaçant avant qu'il ait été sélectionné dans l'échantillon de l'IPC. L'estimation du prix passé, non observé, se fait alors sur la base des évolutions constatées pour des produits similaires. Avec les données de caisse, le prix passé du produit remplaçant, pour peu qu'il ait été vendu, est enregistré dans les données de caisse.

Les données de caisse permettent également l'application des méthodes hédoniques. Ces méthodes reposent sur l'idée que le prix du produit reflète la valorisation des différentes caractéristiques observables des produits. En estimant la dépendance du prix aux caractéristiques observables par une modélisation économétrique, on peut prédire la valorisation de la différence des caractéristiques (« qualité ») en termes de différence de prix. L'utilisation des modèles hédoniques nécessite donc d'une part une connaissance des caractéristiques détaillées des produits et d'autre part un nombre d'observations suffisant pour estimer le modèle économétrique. Les données de caisse assurent le volume des observations et, dans le cas français, le recours à un référentiel d'articles qui décrit chaque code-barres en fonction de caractéristiques permet d'obtenir les variables explicatives du modèle économétrique. Toutefois, maintenir ces modèles économétriques est coûteux en production courante : un modèle doit être développé pour chaque variété de produit et être mis à jour régulièrement. Il paraît difficile de généraliser cette méthode d'estimation à l'ensemble des données de caisse. Elle est utilisée ici dans les tests réalisés à titre de référence.

Sur cinq familles de produits, 6 méthodes d'ajustement qualité ont été proposées :

(1) considérer les produits comme équivalents en termes de qualité et de caractéristiques ; dans ce cas, la différence de prix entre le produit à remplacer observé au mois m et le produit remplaçant observé en $m + 1$ est interprétée comme une pure évolution de prix sans différence de qualité ;

(2) considérer les produits comme des dissemblables purs ; dans ce cas, la différence de prix entre le produit à remplacer observé au mois m et le produit remplaçant observé en $m + 1$ est interprétée comme une pure différence de qualité ;

(3) considérer les produits comme des produits dissemblables en termes de caractéristiques et de qualité, mais corriger la différence de prix entre le produit à remplacer observé au mois m et le produit remplaçant observé en $m + 1$ en considérant que le prix du produit remplacé aurait évolué entre m et $m + 1$ comme les prix observés pour des produits semblables (méthode actuellement utilisée dans l'IPC) ;

(4) considérer les produits comme des produits dissemblables et estimer la différence de qualité comme la différence de prix observée au cours du mois précédent la disparition du produit ;

(5) considérer les produits comme des produits dissemblables et estimer la différence de qualité comme la différence de prix observée deux mois avant la disparition du produit ;

(6) estimer la différence de qualité des deux produits à l'aide d'un modèle hédonique[6].

Les résultats des simulations réalisées (tableaux 3 et 4) montrent que si les coefficients qualité estimés à l'aide de ces différentes méthodes peuvent être légèrement significativement différents du coefficient qualité mesuré avec le modèle hédonique, en revanche, les indices calculés sur la base de ces coefficients ne sont pas significativement différents de ceux calculés à partir d'un modèle hédonique à l'exception de la méthode (1) où aucun ajustement qualité n'est en réalité effectué[7]. Les résultats montrent également que l'algorithme déterministe et l'algorithme alternatif de sélection du remplaçant conduisent à sélectionner des produits différents au point que les indices sans correction de qualité diffèrent significativement (tableau 3). En revanche, ce n'est pas le cas pour les indices corrigés des effets qualité. En conséquence, pour les cas examinés ici, l'indice de prix ajusté pour la qualité est robuste aux modalités de sélection des produits remplaçants.

Pour des raisons d'implémentation et compte tenu de ces résultats, ce sont l'algorithme alternatif et la méthode de recouvrement à 2 mois qui sont retenus pour l'exploitation des données de caisse (pour une présentation plus en détail de ces résultats voir Léonard *et al*., 2017**)**.

6. Par exemple, pour les yaourts, le modèle hédonique retient les variables explicatives suivantes : l'enseigne, la marque, le type de packaging, le parfum, le fait d'être bio, d'incorporer du bifidus, le pourcentage de matières grasses, le pourcentage de sucre, le volume, etc.

7. Le fait que la différence significative entre les coefficients qualité n'ait pas d'impact sur l'indice est lié à la faible fréquence des remplacements d'une part et à la différence peu importante entre les coefficients qualité d'autre part.

Tableau 3
Comparaison des algorithmes de choix du produit remplaçant et des méthodes d'ajustement qualité pour les yaourts, en 2009

Type d'ajustement-qualité	Glissement annuel moyen		Différence entre les coefficients d'ajustement-qualité estimés à partir du modèle hédonique et à partir des autres méthodes			
	Algorithme déterministe (en %)	Algorithme alternatif (en %)	Moyenne*	Distribution de la différence		
				5e centile	Médiane	95e centile
(1) Équivalent	- 4.14 [- 4.5, - 3.8]	- 3.17 [- 3.6, - 2.7]				
(2) Dissemblable pur	- 3.55 [- 3.9, - 3.3]	- 3.51 [- 3.8, - 3.2]	- 0.006 [- 0.017, 0.003]	- 0.22	0.00	0.17
(3) Dissemblable corrigé	- 3.59 [- 3.9, - 3.3]	- 3.56 [- 3.8, - 3.2]	- 0.010 [- 0.020, - 0.001]	- 0.22	0.00	0.16
(4) Recouvrement à 1 mois	- 3.71 [- 4.0, - 3.4]	- 3.60 [- 3.9, - 3.3]	- 0.016 [- 0.024, - 0.009]	- 0.19	- 0.01	0.12
(5) Recouvrement à 2 mois	- 3.60 [- 3.9, - 3.3]	- 3.51 [- 3.8, - 3.2]	- 0.008 [0.016, - 0.001]	- 0.16	0.00	0.13
(6) Modèle hédonique	- 3.52 [- 3.8, - 3.2]	- 3.52 [- 3.8, - 3.2]				

***** La différence moyenne est la différence constatée en moyenne sur un échantillon, entre le coefficient-qualité mesuré à partir du modèle hédonique et ceux mesurés à partir des autres méthodes d'ajustement de la qualité. Une moyenne négative signifie que le coefficient calculé avec la méthode en question est plus grand que celui calculé à partir du modèle hédonique. L'intervalle de confiance à 95 % (indiqué entre crochets) associé a été calculé à partir des valeurs observées sur 100 échantillons, tirés aléatoirement. Quand l'intervalle ne comprend pas la valeur 0, le coefficient d'ajustement qualité diffère significativement de celui calculé avec le modèle hédonique.
Note : pour calculer un indice, les prix ont d'abord été agrégés par variété et point de vente suivant une formule de Laspeyres géométrique puis ces micro-indices ont été agrégés entre eux par une agrégation de Laspeyres arithmétique (pondérée par les ventes de novembre et décembre 2008).
Champ : la taille de l'échantillon a été fixée à 2 %. Les produits ont été sélectionnés proportionnellement à leurs ventes de novembre et décembre 2008 parmi les produits vendus durant ces deux mois.
Source : échantillon de données de caisse pour 17 familles de produits dans 1 000 hyper et supermarchés.

Tableau 4
Comparaison des méthodes d'ajustement qualité pour 5 familles de produits, en 2009

(En %)

Type d'ajustement-qualité	Yaourts	Tablettes de chocolat	Fromage à pâte persillée	Œufs de poule	Café moulu à caféine
Équivalent	- 4.14 [- 4.5, - 3.8]	1.90 [1.4, 2.5]	2.67 [1.87, 3.47]	- 0.58 [- 1.05, - 0.10]	3.35 [2.87, 3.84]
Dissemblable pur	- 3.55 [- 3.9, - 3.3]	- 0.23 [- 0.5, 0.1]	2.43 [1.74, 3.12]	- 0.76 [- 1.09, - 0.43]	3.03 [2.63, 3.43]
Dissemblable corrigé	- 3.59 [- 3.9, - 3.3]	- 0.24 [- 0.6, 0.1]	2.47 [1.78, 3.17]	- 0.78 [- 1.11, - 0.45]	3.19 [2.76, 3.61]
Recouvrement à 1 mois	- 3.71 [- 4.0, - 3.4]	- 0.23 [- 0.5, 0.1]	2.41 [1.71, 3.11]	- 0.82 [- 1.14, - 0.51]	3.19 [2.78, 3.59]
Recouvrement à 2 mois	- 3.60 [- 3.9, - 3.3]	- 0.35 [- 0.7, 0.0]	2.52 [1.90, 3.14]	- 0.81 [- 1.15, - 0.46]	3.19 [2.70, 3.68]
Modèle hédonique	- 3.52 [- 3.8, - 3.2]	- 0.11 [- 0.4, 0.2]	1.961 [1.38, 2.53]	- 0.80 [- 1.19, - 0.40]	3.85 [3.29, 4.42]

Note : pour calculer un indice, les prix ont d'abord été agrégés par variété et point de vente suivant une formule de Laspeyres géométrique puis ces micro-indices ont été agrégés entre eux par une agrégation de Laspeyres arithmétique (pondérée par les ventes de novembre et décembre 2008). Écart-type calculé par boostrap sur 100 échantillons tirés aléatoirement pour les yaourts, 200 pour les tablettes de chocolat, 30 pour les autres familles. Le choix du remplaçant est fait par l'algorithme déterministe.
Champ : la taille de l'échantillon a été fixée de manière arbitraire à 2 %. Les produits ont été sélectionnés proportionnellement à leurs ventes de novembre et décembre 2008 parmi les produits vendus durant ces deux mois.
Source : échantillon de données de caisse pour 17 familles de produits dans 1 000 hyper et supermarchés.

Les prix pratiqués plutôt que les prix affichés

Les prix collectés actuellement dans les points de vente pour calculer l'IPC sont les prix affichés en magasin. Les prix fournis par les données de caisse sont les prix effectivement payés par le consommateur lors du passage en caisse. Ces deux prix peuvent différer par erreur d'affichage de la part du magasin, erreur de relevé lors de la collecte en magasin ou encore en lien avec des promotions réalisées en caisse.

Les organismes internationaux préconisent de suivre les prix réellement pratiqués pour la mesure des indices de prix à la consommation. L'utilisation des données de caisse permet donc de mieux suivre ce que l'on souhaite mesurer. Mais il est toutefois indispensable, pour disposer du prix d'un produit, qu'au moins une vente soit réalisée dans le mois : en l'absence de passage en caisse, aucun prix n'est enregistré alors que le produit peut être proposé à la vente.

Une expérimentation a été réalisée en juin 2014 destinée à comparer les prix figurant dans les bases des données de caisse à des prix affichés, relevés en magasin par les enquêteurs de l'IPC sur la base du code-barres également relevé par les enquêteurs. Pour certains produits de l'IPC, notamment dans les secteurs de l'habillement et des biens durables, aucune vente n'a été trouvée dans les données de caisse. En dehors de ces produits, lorsqu'il existe une vente le jour de la collecte terrain du prix, 90 % des prix sont identiques entre collecte terrain et données de caisse (tableau 5).

Des nouvelles difficultés à traiter

Le GTIN est-il le bon identifiant pour classer les produits ?

L'IPC est un indice à panier fixe. Pour s'assurer que l'on suit le même produit, il faut être en mesure de l'identifier. Actuellement, c'est l'enquêteur qui s'assure de ce suivi en s'appuyant sur la description du produit qu'il relève.

Dans les données de caisse, cette identification doit être automatique : l'intuition suggère qu'elle s'appuie directement sur le code-barres (ou GTIN). Néanmoins, avoir une définition trop stricte de la notion de produit peut conduire à masquer des évolutions de prix. C'est le problème que soulève l'utilisation directe du GTIN pour définir le produit suivi dans l'IPC. En effet, plusieurs codes-barres peuvent être utilisés pour identifier un même produit pour le consommateur et donc au sens de l'IPC. Différents exemples de ce phénomène ont pu être constatés : 1) des produits identiques sont fabriqués dans différentes usines et les fabricants utilisent différents codes-barres pour identifier l'unité de production du produit ; 2) le code-barres est modifié lors de relances commerciales. Lors de ces relances, une modification du packaging, en général sans impact sur l'utilité du consommateur, s'accompagne éventuellement d'un changement de prix. Correspondant à des processus de fabrication différents, les codes-barres sont modifiés ; 3) cas similaire à la relance commerciale, mais temporaire, la promotion fabricant correspond, par exemple, à des cadeaux offerts avec un produit (e.g. un verre avec une bouteille de vodka), des bons de réduction attachés au produit, des conditionnements exceptionnels, ou encore des quantités offertes. Toutes ces promotions impliquent une modification du procédé de fabrication du produit fini et ce faisant, des codes-barres associés.

Considérer qu'une promotion ou une relance est un produit différent n'est pas sans conséquence sur la mesure de l'évolution des prix. La baisse ou la hausse de prix liée à la promotion ou à la relance ne seraient en effet pas prises en compte. Même dans le cas où le produit initial disparaît complètement et est remplacé par sa relance/promotion, les traitements qualité mis en place lors du remplacement, par recouvrement, annulent tout effet sur les prix.

Tableau 5
Comparaison des prix des données de caisse (DDC) et des prix collectés sur le terrain, en nombre de relevés, en juin 2014

	Secteurs de consommation				Ensemble
	Alimentation	Biens durables	Habillement	Biens manufacturés	
Ensemble des relevés	526	65	128	234	953
dont :					
pas de vente dans DDC le jour du relevé	20 %	89 %	90 %	63 %	44 %
prix identique dans DDC et relevé enquêteur	72 %	9 %	6 %	35 %	50 %
prix différent en défaveur du consommateur	4 %	0 %	0 %	2 %	2 %

Note : 526 prix ont été comparés pour des produits alimentaires. Pour 20 % des relevés, aucun prix n'était disponible dans les données de caisse le jour donné, faute de ventes ; dans 72 % des cas, le prix était identique.
Champ : 953 relevés utilisés dans l'IPC en juin 2014 et les prix correspondant dans les données de caisse.
Source : Insee, IPC, données de caisse.

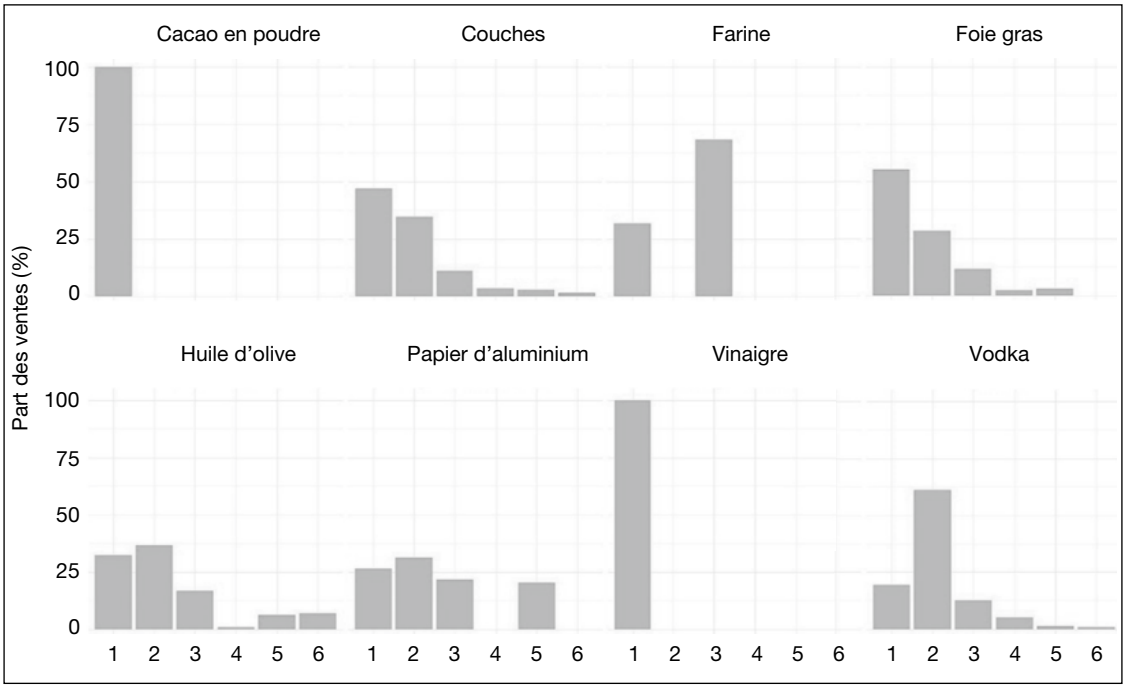
Afin de mesurer correctement les évolutions de prix, en prenant en compte ces éventuelles relances ou promotion, le panier de produits n'est pas constitué de codes-barres mais de « classes d'équivalence », des regroupements de codes-barres pour lesquels on considère que le produit est identique aux yeux du consommateur. Reste à définir ce qu'est un produit identique aux yeux du consommateur. L'usage consiste à considérer que si des modifications ténues du produit suivi n'apportent pas de modification notable de l'utilité du consommateur, alors le produit reste le même. Ces modifications peuvent porter sur l'emballage (sans changement de contenu), sur les quantités vendues[8] pourvu que les modifications demeurent dans une fourchette proche (fixée conventionnellement de 1 à 2 dans l'IPC) ou toute autre caractéristique qui n'altère pas la nature du produit.

Pour définir un produit identique avec les données de caisse, on s'appuie sur un référentiel d'articles qui décrit chaque code-barres en fonction d'un certain nombre de caractéristiques. Ces caractéristiques doivent être identiques, à l'exception du volume de produit qui peut changer dans une proportion plus ou moins importante. Parmi ces caractéristiques, différentes pour chaque famille (entre 10 et 30 caractéristiques selon les familles), on peut citer la marque, la quantité vendue, l'emballage, le parfum, le taux de matière grasse, le fait d'être bio ou non, etc. À titre d'exemple, les codes-barres de 8 variétés ont été regroupés en classes d'équivalence sur les années 2013 à 2015. Sur ces huit variétés, le nombre maximal de codes-barres par classe d'équivalence est très faible (en l'occurrence, 6) et à l'exception d'une ou deux variétés, la part des ventes associée à des classes d'équivalence contenant plus d'un code-barres est, sauf exception, inférieure à 10 % (figure III).

Calculer un indice à partir de différents codes-barres nécessite d'agréger plusieurs codes-barres par classe d'équivalence, sur un mois et dans un point de vente donnés. Les produits composant une classe d'équivalence étant par définition homogènes et en accord avec la pratique recommandée au niveau international pour traiter les promotions, les prix des différents codes-barres sont agrégés en calculant une valeur unitaire, le prix suivi étant un prix rapporté à une unité de volume ou de masse.

8. Le prix suivi est, systématiquement dans l'IPC, un prix rapporté à une unité de volume ou de masse.

Figure III
Nombre de codes-barres par classes d'équivalence pour quelques variétés sur la période 2013-2015



Note : pour la variété « couches », les classes d'équivalence composées d'un seul code-barres représentent près de 50 % des ventes. Pour la même variété, approximativement 30 % des classes d'équivalence comprennent deux codes-barres.
Champ : prix des produits représentant les 8 variétés présentées.
Source : données de caisse de 4 enseignes représentant 30 % du marché, de 2013 à 2015.

Classer les produits dans la nomenclature, une tâche gigantesque

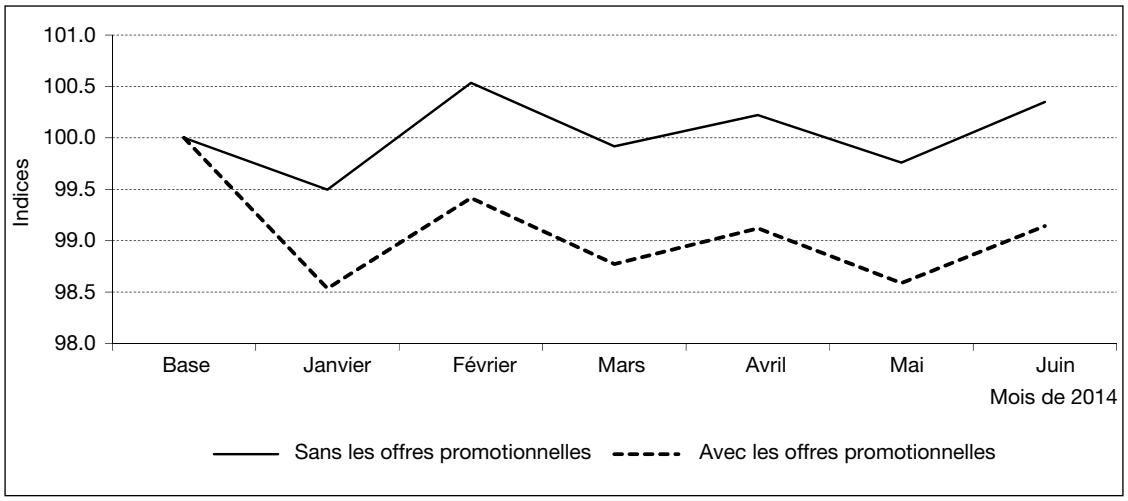
Une fois les produits identifiés par classe d'équivalence grâce à une combinaison du code-barres et du référentiel des articles, reste encore la tâche de classer les produits par variétés puis dans la nomenclature de fonction de consommation. Cette tâche est nécessaire d'une part pour des problèmes de diffusion et de statistiques produites : l'IPC est actuellement diffusé selon la nomenclature COICOP (*Classification of Individual Consumption by Purpose*) correspondant à une partition élémentaire de la consommation en 303 postes. Il convient donc de classer les codes-barres selon une nomenclature relativement détaillée de produits (par exemple, plats cuisinés à base de viande, huile d'olive, etc.). Il existe un niveau plus fin encore, la variété, qui définit le périmètre sur lequel on effectue les hypothèses de substituabilité déjà évoquées. Dans l'approche classique où environ un millier de variétés sont suivies, c'est l'enquêteur qui classe le produit par variété. Le recours aux données de caisse, exhaustives sur leur champ, rend ce classement manuel impossible. Pour la plupart des autres pays, c'est une des principales difficultés des données de caisse car ils ne disposent pas d'un référentiel des articles. La classification des produits se fait alors sur la base de la description par enseigne des produits, parfois sommaire, et qui nécessite souvent d'avoir recours à des outils de *machine learning*. Dans le cas français, l'existence d'un référentiel des articles permet de classer ces données, très volumineuses mais relativement structurées, à l'aide d'une simple table de passage du référentiel à une nomenclature de fonction. La difficulté tient en réalité à la définition des variétés elles-mêmes.

Si la nomenclature de fonction est relativement fine et constitue une partition de la consommation, les variétés sont conçues, dans l'approche classique comme des « représentants » du niveau le plus fin de la nomenclature de fonction et n'ont pas vocation à former une partition de la consommation. Par exemple, le poste huile d'olive sera représenté par une unique variété : une huile avec un volume dans une fourchette définie, un niveau de raffinement défini, un emballage en verre. Ces variétés sont définies à dire d'experts. Avec les données de caisse et la volonté de les exploiter dans leur intégralité, la définition des variétés doit être, sinon automatisée, du moins fortement assistée pour permettre aux experts de traiter convenablement une masse conséquente d'informations.

De nouveaux phénomènes : les produits saisonniers

La connaissance exhaustive de la consommation des ménages fait apparaître de nouveaux phénomènes, qui, s'ils ne sont pas traités de manière appropriée, peuvent biaiser l'IPC. Les produits saisonniers en sont un exemple. La saisonnalité des produits n'est pas, en soi, un problème nouveau pour l'IPC : l'observation sur une période seulement de l'année de certains produits amène, afin de rester représentatif de l'ensemble de la consommation des ménages, y compris des produits saisonniers, à imputer les prix en l'absence saisonnière du produit. Actuellement, le champ des produits saisonniers est bien défini : certains fruits et légumes, des vêtements, certains services (par exemple, les remontées mécaniques ou les emplacements de camping) ne sont observables qu'une partie de l'année. La nouveauté avec les données de caisse est la généralisation de ces produits saisonniers, non suivis jusqu'à présent car les enquêteurs ont la consigne de ne suivre que des produits dits « bien suivis et bien vendus » et excluent de leur sélection les produits éphémères. Les chocolats de Pâques, les conditionnements pour Noël, certaines glaces disponibles uniquement l'été ne sont ainsi pas suivis. La difficulté tient à identifier ces saisonnalités afin de les traiter comme telles. Ne pas comprendre qu'un produit est saisonnier et le traiter comme un produit classique, c'est-à-dire disparaissant et étant remplacé par un autre à l'aide d'un ajustement qualité, peut engendrer de fortes dérives de l'indice. Un cas emblématique est celui des saumons fumés dont les grands conditionnements vendus uniquement en période de fin d'année génèrent un chiffre d'affaires conséquent. Présents en décembre, ils sont en promotion début janvier et ont disparu des rayons début février. Si ces conditionnements ne sont pas identifiés comme une variété saisonnière, ils sont remplacés en février par un paquet de plus petite taille, avec un ajustement qualité par recouvrement, et la baisse temporaire de prix observée en janvier et liée aux promotions sur les gros conditionnements est définitivement enregistrée dans l'indice, y compris pour les plus petits conditionnements alors qu'elle ne les concerne pas (figure IV).

Figure IV
Indices des produits de la famille des poissons fumés au rayon frais avec et sans offres promotionnelles, base 100 en décembre 2013

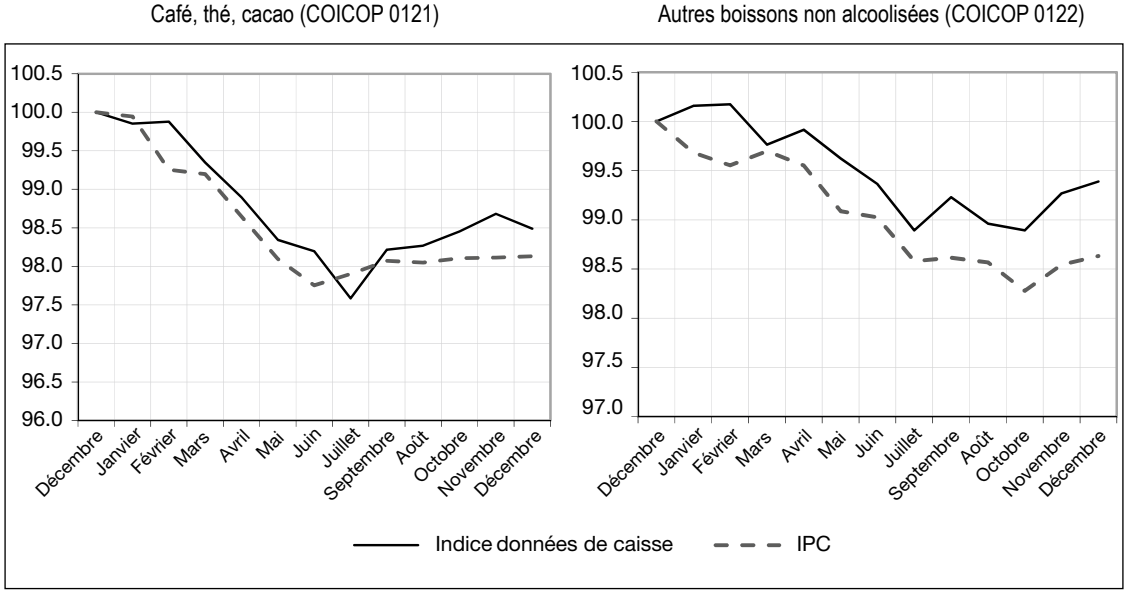


Note : en prenant en compte les offres promotionnelles, l'indice des prix des poissons fumés a chuté de 1.5 % en janvier 2014.
Champ : poissons fumés au rayon frais.
Source : données de caisse de 4 enseignes représentant 30 % du marché, en 2014.

* *
*

Sur la base de la méthodologie définie dans cet article, de premiers indices ont pu être construits sur l'ensemble du champ de l'alimentation industrielle. Ils montrent que données de caisse et collecte terrain permettent d'approcher une mesure globalement similaire de l'inflation pour les postes comparables, c'est-à-dire où les produits sont principalement vendus en super et hypermarchés (figure V). Sur la base de ces

Figure V
Indices des prix à la consommation pour deux postes et indices calculés uniquement sur le champ données de caisse en 2014, base 100 en décembre 2013



Champ : pour l'IPC, toutes formes de ventes ; pour les données de caisse, super et hyper marché ; hors promotion pour les données de caisse.
Source : IPC, données de caisse de 4 enseignes représentant 30 % du marché.

études, les données de caisse, dont la transmission par les enseignes est dorénavant obligatoire (encadré 5), seront utilisées pour produire l'IPC publié mensuellement par l'Insee, à l'horizon de 2020, après une année de répétition générale en 2019. À terme, les données de caisse devraient permettre de répondre à des demandes nouvelles : indices régionaux sur des champs restreints, comparaison spatiale de niveau de prix (voir par exemple Léonard *et al*., ce numéro), indices de prix pour des micro-segments de consommation. □

ENCADRÉ 5 – **L'obtention des données de caisse, un nouveau cadre législatif français**

En France, la production de statistiques et notamment la production d'enquêtes est encadrée par la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques. Certaines enquêtes, jugées d'intérêt public, peuvent être obligatoires, par arrêté du ministre chargé de l'économie. L'exploitation, à des fins d'information générale, de données collectées par des administrations, des organismes publics ou des organismes privés chargés d'une mission de service public, est également prévue et définie.

En revanche, le recours à des données privées pour des fins statistiques n'était pas prévu, jusqu'à la loi du 7 octobre 2016 pour une République numérique, et la transmission de telles données, actifs privés des entreprises, ne pouvait être obligatoire. Dans le même temps, un certain nombre de ces données privées apparaissaient comme de nouvelles sources prometteuses pour la statistique : données de caisse mais également données issues de la gestion des opérateurs de téléphonie mobile, de la gestion des transactions de cartes bancaires, des sites d'offres d'emploi, etc.

Afin d'encadrer le recours à de telles données, la loi pour une République numérique prévoit que le ministre chargé de l'économie peut décider, après avis du Conseil national de l'information statistique (CNIS), que les personnes morales de droit privé sollicitées pour des enquêtes transmettent par voie électronique sécurisée au service statistique public, à des fins exclusives d'établissement de statistiques, les informations présentes dans les bases de données qu'elles détiennent, lorsque ces informations sont recherchées pour les besoins d'enquêtes statistiques obligatoires.

Depuis le 13 avril 2017, un arrêté signé par le ministre de l'économie, rend obligatoire la transmission des données de caisse par les commerces de détail en magasin non spécialisés à prédominance alimentaire de plus de 400m^2. Il fiabilise et garantit ainsi l'accès aux données de caisse, un préalable lorsque l'on veut construire un indice, l'IPC, produit dans des délais très courts et non révisable.

BIBLIOGRAPHIE

Chessa, A. (2015). Towards a generic price index method for scanner data in the Dutch CPI. Paper for the fourteenth Ottawa Group Meeting. https://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s1room2.pdf

Diewert, E., Fox, K. & Ivancic, L. (2009). Scanner Data, Time Aggregation and the Construction of Price Indexes. Paper for the eleventh Ottawa Group Meeting. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1bd88ae9af79cfa1ca257693001bb7fa/$FILE/2009 11th meeting - Lorraine Ivancic kevin Fox (University of New South Wales) and W. Erwin Diewert (University of British Columbia)_Scanner Data Time Agg.pdf

Diewert, E. & Fox, K. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. Paper for the fifteenth Ottawa Group Meeting. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Substitution bias in multilateral methods for CPI construction using scanner data -Erwin Diewert, Kevin Fox -Paper.pdf

Eurostat (2013). *Compendium of HICP reference documents.* https://ec.europa.eu/eurostat/documents/3859598/5926625/KS-RA-13-017-EN.PDF/59eb2c1c-da1f-472c-b191-3d0c76521f9b

Eurostat (2017). *Practical Guide for Processing Supermarket Scanner Data.* 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.* https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331155.pdf

Jaluzot, L. & Sillard, P. (2016). Échantillonnage des agglomérations de l'IPC pour la base 2015. Insee, *Document de travail* N° F1601.
https://www.insee.fr/fr/statistiques/2022137

Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2017). Données de caisse et ajustements qualité. Insee, *Document de travail* N° F1704.
https://www.insee.fr/fr/statistiques/2912650

Léonard, I., Sillard, P. & Varlet, G. (2019). Écarts spatiaux de prix dans l'alimentaire avec les données de caisse. *Economie et Statistique / Economics and Statitistics,* ce numéro.

Sillard, P. (2017). Indices des prix à la consommation. Insee, *Document de travail* N° F1706.
https://www.insee.fr/fr/statistiques/2964204

Von der Lippe, P. (2012). Notes on GEKS and RGEKS indices – Comments on a method to generate transitive indices. *Munich Personal RePEc Archive.*
http://www.von-der-lippe.org/dokumente/MPRA_paper_42730.pdf

Zhang, L. C., Johansen, I. & Nygaard, R. (2017). Testing unit value data price indices. Paper for the fifteenth Ottawa Group Meeting.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Testing unit value data price indices - Li-Chun Zhang, Ingvild Johansen, Ragnhild Nygaard - Paper.pdf