N° 509 - 2019

Economie ET Statistique

Economics AND Statistics

Big Data et statistiques
2ème partie
Les Big Data dans l'indice des prix à la consommation

Big Data and Statistics
Part 2
Big Data in the Consumer Price Index

Insee

Economie ET Statistique

Economics AND Statistics

Directeur de la publication / Director of Publication:
Jean-Luc TAVERNIER
Rédactrice en chef / Editor in Chief:
Sophie PONTHIEUX
Responsable éditorial / Editorial Manager: Pascal GODEFROY
Assistant éditorial / Editorial Assistant: Étienne de LATUDE
Traductions / Translations: RWS Language Solutions
Chiltern Park, Chalfont St. Peter, Bucks, SL9 9FG Royaume-Uni
Maquette PAO et impression / CAP and printing: JOUVE
1, rue du Docteur-Sauvé, BP3, 53101 Mayenne

Conseil scientifique / Scientific Committee

Jacques LE CACHEUX, président (Université de Pau et des pays de l'Adour)
Jérôme BOURDIEU (École d'économie de Paris)
Pierre CAHUC (Sciences Po)
Gilbert CETTE (Banque de France et École d'économie d'Aix-Marseille)
Yannick L'HORTY (Université de Paris-Est - Marne la Vallée)
Daniel OESCH (Life Course and Inequality Research (LINES) et Institut des sciences sociales - Université de Lausanne)
Sophie PONTHIEUX (Insee)
Katheline SCHUBERT (École d'économie de Paris, Université Paris I)
Claudia SENIK (Université Paris-Sorbonne et École d'économie de Paris)
Louis-André VALLET (Observatoire sociologique du changement-Sciences Po/CNRS)
François-Charles WOLFF (Université de Nantes)

Comité éditorial / Editorial Advisory Board

Luc ARRONDEL (École d'économie de Paris)
Lucio BACCARO (Max Planck Institute for the Study of Societies-Cologne et Département de Sociologie-Université de Genève)
Antoine BOZIO (Institut des politiques publiques/École d'économie de Paris)
Clément CARBONNIER (Théma/Université de Cergy-Pontoise et LIEPP-Sciences Po)
Erwan GAUTIER (Banque de France et Université de Nantes)
Pauline GIVORD (Ocde et Crest)
Florence JUSOT (Université Paris-Dauphine, Leda-Legos et Irdes)
François LEGENDRE (Erudite/Université Paris-Est)
Claire LELARGE (Université de Paris-Sud, Paris-Saclay et Crest)
Claire LOUPIAS (Direction générale du Trésor)
Pierre PORA (Insee)
Ariell RESHEF (École d'économie de Paris, Centre d'économie de la Sorbonne et CEPII)
Thepthida SOPRASEUTH (Théma/Université de Cergy-Pontoise)

OÙ SE PROCURER
Economie et Statistique / Economics and Statistics

Les numéros sont en accès libre sur le site www.insee.fr. Il est possible de s'abonner aux avis de parution sur le site.

La revue peut être achetée sur le site **www.insee.fr** via la rubrique « Acheter nos publications ». La revue est également en vente dans 200 librairies à Paris et en province.

WHERE TO GET
Economie et Statistique / Economics and Statistics

All the issues and articles are available in open access on the Insee website **www.insee.fr.** Publication alerts can be subscribed on-line.

The printed edition of the journal (in French) can be purchased on the Insee website **www.insee.fr** and in 200 bookshops in Paris and province.

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Directeur Général : Jean-Luc TAVERNIER
Direction Générale : 88, avenue Verdier, CS 70058, 92541 MONTROUGE Cedex
Tél : +33 (0)1 87 69 50 00

N° 509 - 2019

Economie ET Statistique

Economics AND Statistics

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes,
et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

Economie et Statistique / Economics and Statistics

Numéro 509 – 2019

BIG DATA ET STATISTIQUES
2ème partie

LES BIG DATA DANS L'INDICE DES PRIX À LA CONSOMMATION

Introduction – La chaîne de valeur des données de caisse et des données moissonnées sur le Web

Introduction – The Value Chain of Scanner and Web Scraped Data

Jens Mehrhoff*

Résumé – Avec l'avènement des données de caisse et des données du Web, les « big data » trouvent de plus en plus leur place dans les statistiques officielles. Cette deuxième partie du numéro spécial « Big Data et statistiques » est consacrée à l'évolution de l'utilisation de ces données pour les indices des prix à la consommation. Dans quelle mesure les données massives sont-elles différentes des données de sources plus traditionnelles, comme la collecte des prix sur le terrain, et comment changent-elles le processus de production des indices des prix à la consommation ? Les quatre articles de ce numéro spécial traitent de ces questions à partir de l'expérience acquise par les instituts de statistique de la France, de la Suède et des Pays-Bas. Cette introduction les met en perspective par rapport à la chaîne de valeur des données de caisse et des données moissonnées sur le Web et évoque quelques autres enjeux pour la recherche dans ce domaine.

Abstract *– With the advent of scanner and web scraped data, "big data" sources are increasingly finding their way into official statistics. This second part of the special issue on "Big Data and Statistics" is devoted to developments in the use of these data for consumer price indices. To what extent are big data different to more traditional data sources such as the collection of prices in the field, and how do they change the process of producing consumer price indices? The four papers in this special issue address these questions by means of the experiences gained in the statistical offices of France, Sweden and the Netherlands. This introduction puts them into perspective vis-à-vis the value chain of scanner and web scraped data and looks at some further issues for research in this field.*

Codes JEL / JEL Classification : C43, C55, C82, E31
Mots-clés : indice des prix à la consommation, IPC, big data, données de caisse, données moissonnées
Keywords: consumer price indices, big data, scanner data, web scraped data

Rappel :

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes, et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

* *Deutsche Bundesbank (jens.mehrhoff@bundesbank.de)*

Reçu le 21 juillet 2019
Traduit de la version originale en anglais

Mehrhoff, J. (2019). Introduction – The Value Chain of Scanner and Web Scraped Data. *Economie et Statistique / Economics and Statistics*, 509, 5–11.
https://doi.org/10.24187/ecostat.2019.509.1980

Le contexte

Les indices des prix à la consommation sont la référence pour évaluer la stabilité des prix, ce qui en fait les indicateurs les plus importants pour la définition des politiques monétaires par les banques centrales. Avec l'arrivée des données de caisse et des données moissonnées sur le Web, les sources de « données massives » prennent de plus en plus d'importance dans la production des indices des prix à la consommation, et ce à l'échelle mondiale. Cette seconde partie du numéro spécial « Big Data et statistiques » est consacrée aux développements de l'utilisation des données de caisse et des données moissonnées pour l'élaboration des indices des prix à la consommation.

Les quatre articles de ce numéro spécial soulèvent deux questions sous-jacentes. Premièrement, dans quelle mesure les données massives sont-elles différentes des sources de données classiques telles que la collecte de prix sur le terrain, ou leur ressemblent-elles ? Deuxièmement, comment ces données massives modifient-elles le processus de production des indices des prix à la consommation ? Si l'objectif est le même quelle que soit la source de données, à savoir mesurer le taux de variation des prix à la consommation, la façon dont ce chiffre est obtenu peut varier. Tout d'abord, les données de caisse et les données moissonnées sur le Web permettent d'accéder à un ensemble de produits beaucoup plus large que l'échantillonnage classique. Cette couverture des biens et services est en principe supérieure, certes, mais est également plus chère en raison du renouvellement dû aux produits nouveaux ou sortants – en d'autres termes, l'univers des produits est dynamique. En outre, il est également possible d'obtenir de l'information sur les quantités vendues (avec les données de caisse), ou au moins un classement des articles par popularité (avec les données moissonnées sur le Web), ce qui permet de calculer des indices pondérés plutôt que de devoir se fier à des formules non pondérées. Ici, le prix à payer est la « dérive de chaîne », c'est-à-dire que l'indice peut faire apparaître des tendances erronées au fil du temps.

Dans cette introduction, nous plaçons les quatre articles de ce numéro dans le contexte de la chaîne de valeur des données de caisse et des données du Web, tenant compte de trois phases stylisées : i) la collecte des données ; ii) le traitement des données ; et iii) la diffusion des résultats. Nous concluons sur quelques perspectives de recherche supplémentaires dans ce domaine.

La collecte des données

Grâce aux pionniers de l'utilisation de ces nouvelles sources de données, les bonnes pratiques en matière de collecte de données de caisse et de données moissonnées sont maintenant connues. Le *Practical Guide for Processing Supermarket Scanner Data* (guide pratique du traitement des données de caisse des supermarchés) publié par Eurostat en 2017 fournit des recommandations qui, de façon générique, s'appliquent également en dehors du contexte des données de caisse des supermarchés. Un point qui apparaît essentiel est d'établir une relation avec les propriétaires des données. Les chaînes de supermarchés et les magasins en ligne craignaient que leurs données ne soient utilisées de façon abusive par leurs concurrents, craintes qui disparaissent une fois qu'une relation de confiance est instaurée avec les instituts de statistique.

Pour les données de caisse, il est possible de mettre en place une forme de contrepartie, c'est-à-dire que les fournisseurs des données reçoivent des indices de référence du marché, ainsi que des analyses de données, en échange des chiffres qu'ils communiquent. En aucun cas il ne s'agit de diffuser les micro-données ou des informations sur les concurrents. S'agissant des données moissonnées sur le Web, le propriétaire du site, s'il sait qui utilise ses données et dans quel but, peut être disposé à fournir une interface de programme d'application (API) plutôt que de bloquer l'adresse IP de l'institut de statistique.

Une autre démarche pour la collecte des données consiste à établir un cadre juridique permettant aux instituts de statistique d'accéder à ces sources ; les modalités précises dépendent en grande partie des dispositions institutionnelles en vigueur au niveau national.

Quel que soit le niveau d'agrégation souhaité ou possible en termes de durée, de magasins et de régions, des jeux de données expérimentales devraient être testés avant d'intégrer les flux de données dans la chaine de production. Des deux côtés, cela implique d'avoir résolu de nombreux problèmes techniques comme le format de diffusion ou le stockage des données.

Le traitement des données

Il y a eu plusieurs approches pour décomposer la phase de traitement des données de façon plus précise. Bien qu'elles soient globalement semblables, certains aspects sont néanmoins différents en raison de dispositions institutionnelles qui peuvent être particulières à un institut de statistique en matière de prix à la consommation. Les étapes courantes incluent, sans s'y limiter, la classification automatique des produits, l'agrégation intermédiaire des produits « homogènes », le filtrage des observations en fonction de règle spécifiques et le calcul de l'indice définitif.

Dans ce registre, **Marie Leclair et ses co-auteurs** examinent comment un certain nombre de questions ont été traitées en France pour l'agrégation des prix dans la production des indices, le traitement des ajustements de la qualité, le classement des produits par variété homogène et le traitement des relances et des promotions.

Classification

Compte tenu des vastes volumes de produits couverts par les données massives, il n'est plus possible de les classer dans la nomenclature COICOP (ou dans des subdivisions de cette nomenclature) de façon manuelle, cela ne peut se faire que de façon automatique. Le classement peut être établi par le propriétaire des données, au moins en partie. Les supermarchés, par exemple, ont établi leur propre classification pour les données de caisse, qui pourrait être utile à un classement automatique. Il en est de même pour les magasins en ligne, où les produits peuvent être présentés de façon structurée. Toutefois, si les informations sur ces classifications ne sont pas disponibles ou ne sont pas suffisamment détaillées, il faut alors avoir recours à des techniques d'apprentissage automatique supervisé. Cela implique alors de construire un petit jeu de données labellisées afin d'entraîner l'algorithme.

Pour commencer, tous les produits doivent être classés. En plus des informations fournies par le propriétaire des données, les codes de produits (comme les GTIN), les descriptions (texte) et d'autres métadonnées (comme la taille) sont habituellement disponibles. À cet égard, « l'ingénierie des caractéristiques » *(feature engineering)* pose un problème majeur. Dans la plupart des cas, les descriptions de produits ne sont pas du texte standard mais utilisent un vocabulaire particulier et différents types d'abréviations. En règle générale, les codes de produits suivent un certain type de structure. En outre, chaque mois, de nouveaux produits apparaissent et doivent également être classifiés. Les produits déjà classés ne devraient pas être reclassés dans le cadre de cet exercice. Quoi qu'il en soit, la qualité du classement au fil du temps devrait être évaluée. Une autre difficulté est liée à l'identification des relances, par exemple lorsque le même produit est vendu dans un emballage différent et reçoit un nouveau code.

Agrégation des produits

Pour calculer les indices élémentaires, la première étape consiste à définir le produit dit « homogène ». En raison du taux de renouvellement des produits et du volume significatif

des observations, l'approche classique du panier fixe n'est viable que si un échantillon de petite taille mais fixe est tiré des données. Avec l'approche consistant à utiliser la plupart des données collectées, un compromis doit être trouvé entre l'homogénéité et la continuité du produit, problème qui est accentué par les relances, qui ne sont pas faciles à identifier.

Ici, le dilemme vient du fait que, par définition, il n'y a pas de solution optimale. Il est judicieux, d'une part, de tester des scénarios variés pour la définition du produit et, d'autre part, d'analyser indépendamment une mesure d'homogénéité et une mesure de continuité, ainsi que leur évolution au fil du temps, plutôt qu'une seule statistique synthétique. Les produits présentant un taux de renouvellement élevé et les produits saisonniers requièrent une attention particulière. Dans le secteur de l'électronique grand public, par exemple, un ajustement hédonique de la qualité pourrait être la meilleure solution. Au bout du compte, la continuité du produit ne doit pas être acquise aux dépens d'un biais (de la valeur unitaire).

À titre d'illustration de ces difficultés, l'article de **Can Tongur** traite de la préservation de l'approche du panier fixe, en dépit de l'introduction des données de caisse en Suède, et cherche à évaluer si la méthode classique de remplacement manuel d'articles, accompagnée d'ajustements de la qualité et de la quantité, reste pertinente pour assurer la comparabilité au fil du temps et entre pays.

Filtrage

Si un échantillon fixe est tiré des données, les problèmes associés aux données de caisse et aux données moissonnées sur le Web sont semblables à ceux qui se posent dans le cadre de la collecte de prix traditionnelle, notamment en termes d'imputation et d'ajustement de la qualité. Si le but est d'utiliser la plupart des informations disponibles, en revanche, certaines règles sont nécessaires pour pré-traiter les données brutes. Les filtres suppriment habituellement les codes de produits qui ne sont pas représentatifs au fil du temps, les observations jugées suspectes et, éventuellement, les produits dont les ventes sont faibles ou qui sont susceptibles d'être retirés de la vente.

Les codes de produits non représentatifs incluent les groupes de produits hors du champ (par exemple des vêtements pour les supermarchés) et les codes génériques utilisés par le propriétaire des données d'une façon non stable. Les observations suspectes renvoient à la fois aux valeurs aberrantes (prix exceptionnellement bas ou erronés) et aux produits influents (c'est-à-dire une part des dépenses extrême ou un effet de levier important). Le filtre visant à identifier les ventes faibles produit une pondération grossière, ne laissant que les produits pertinents au sein de l'indice et imitant ainsi une formule pondérée. Les filtres identifiant les produits liquidés tentent de minimiser l'effet de baisse des produits sortants lors des ventes de liquidation.

Calcul de l'indice

Une fois que le jeu de données a été éventuellement retravaillé, par exemple pour l'imputation des prix manquants, l'indice définitif peut être calculé. Les options incluent un panier fixe avec une formule bilatérale, ou des approches multilatérales avec un univers de produits dynamique. En aucun cas les indices pondérés ne devraient être chaînés à une fréquence élevée (par exemple mensuelle), sous peine d'un risque sévère de dérive de l'indice.

Si une approche bilatérale est choisie, on se retrouve dans la même situation qu'avec la collecte de prix classique. La différence principale repose sur le fait que, si des données de caisse sont utilisées, il est possible d'utiliser les poids de la période actuelle et des formules telles que celles de Fisher ou Tornqvist. En revanche, si une approche multilatérale est choisie, plusieurs décisions doivent être prises : quelle approche multilatérale spécifique

faut-il mettre en œuvre, avec combien de mois pour la fenêtre d'estimation, et comment les séries chronologiques diffusées peuvent être étendues en temps réel sans révision. Il n'y a pas de « bonne réponse » consensuelle à ces questions, et il pourrait être plus simple de chercher des méthodes robustes – qui permettent d'établir des estimations fiables même pour les groupes de produits difficiles – plutôt que des justifications économiques ou statistiques.

Bien qu'elles soient désormais utilisées dans les comparaisons intertemporelles, les approches multilatérales trouvent leur origine de travaux sur les comparaisons internationales des parités de pouvoir d'achat. Si ces approches ne sont donc pas adaptées spécifiquement au problème du calcul d'indices, elles font néanmoins l'affaire en permettant d'éviter toute dérive de chaîne, ce qui est absolument essentiel. De nombreuses méthodes ont été suggérées pour les comparaisons interspatiales, mais les trois approches suivantes sont préférables dans le domaine temporel (citées ici sans ordre particulier) : l'indicatrice temps/produit, la méthode Geary-Khamis et la méthode Gini-Eltetö-Köves-Szulc. L'indicatrice temps/produit établit l'indice de prix dans un cadre de régression log-linéaire, la méthode Geary-Khamis le fait à travers les valeurs propres d'une fonction harmonique et la méthode Gini-Eltetö-Köves-Szulc transitivise des indices bilatéraux par le biais d'une moyenne géométrique.

Bien que ces trois approches satisfassent l'exigence de circularité, c'est-à-dire que l'indice chaîné défini comme le produit des indices à court terme doit être égal à l'indice direct, l'ensemble de la série doit être révisée lorsque les données du mois suivant sont ajoutées. C'est malheureusement inévitable, quelle que soit la méthode choisie. Pour contourner ce problème des révisions, la fenêtre d'estimation est avancée tout en maintenant sa durée, et le nouvel indice est raccordé sur un chiffre déjà diffusé. En règle générale, les fenêtres d'estimation devraient couvrir au moins 13 mois et le raccordement devrait être effectué sur le mois précédent (raccordement des variations), sur le même mois de l'année précédente (raccordement des intervalles) ou similaire.

Les articles se multiplient sur la question de la durée de la fenêtre d'estimation, qui pose des problèmes particuliers pour les articles fortement saisonniers présentant des tendances spécifiques, ainsi que sur la façon dont l'extension devrait être effectuée. Dans la mesure où le chaînage des indices des prix à la consommation est aujourd'hui la norme, on pourrait au moins répondre à ce problème en examinant la façon dont l'indice global est calculé. Les résultats suggèrent qu'une forme d'ancrage peut atténuer la dépendance de l'indice à son évolution. Les approches de chaînage classiques reflètent cette situation en faisant référence soit à la moyenne de l'année précédente (recouvrement annuel) soit au dernier trimestre/mois de l'année précédente (recouvrement sur un trimestre/mois).

À cet égard, un dernier mot s'impose. Bien qu'il soit déjà régressif d'inventer encore une autre méthode qui se rapproche de l'indice de référence entièrement transitif avec l'un ou l'autre des jeux de données, des difficultés importantes surviennent avec cet indice de référence, notamment en cas d'absence saisonnière d'un produit. Une extension de l'intervalle de temps a l'effet inverse ; l'indice perd de sa « caractéristicité ». Qu'est-ce que cela signifie ? Les différences relatives constatées entre les niveaux de prix des produits sont prises en compte de façon implicite dans les méthodes multilatérales. Cet ajustement représente une moyenne sur la fenêtre d'estimation. Toutefois, si les produits inclus dans l'agrégat élémentaire présentent des tendances différentes, cette moyenne temporelle est tout simplement erronée (elle n'est pas « stationnaire »). Pour les articles fortement saisonniers tout particulièrement, cela peut engendrer des chiffres imprécis dans l'indice de référence et des fenêtres d'estimation différentes peuvent déboucher sur des séries temporelles largement divergentes. Un exemple de calcul de l'indice se trouve dans l'article d'**Antonio Chessa et Robert Griffioen**. Plus précisément, leur article tente de déterminer si, étant donné la rareté des données de transactions, les prix des biens de consommation moissonnés sur le Web sont une alternative possible aux données de caisse.

Diffusion des résultats

Les instituts de statistique sont peu susceptibles de diffuser des informations très détaillées, encore moins si elles permettent d'identifier le propriétaire des données. Pour cette raison, les indices élémentaires sont agrégés à partir de ce niveau très détaillé, éventuellement au niveau régional, à la nomenclature COICOP, en utilisant des poids issus par exemple de statistiques d'entreprises. Mais cela signifie également que le niveau de détail offert aux utilisateurs des données est souvent le même avec la publication de données de caisse ou de données moissonnées sur le Web qu'avec la collecte de prix classique. En quelque sorte, les instituts de statistique utilisent des sources de « big data » mais ne diffusent que des « petites statistiques ».

En outre, les indices établis à partir de données de caisse et de données moissonnées sont plus volatils que les indices classiques. Alors que la collecte traditionnelle des prix des modèles appariés n'entraine que peu ou pas de bruit dans l'évolution des prix, les nouvelles méthodes introduisent beaucoup de bruit dans les séries temporelles. C'est d'autant plus vrai pour les indices pondérés et l'utilisation de données de caisse. Essentiellement, et malgré la fenêtre d'estimation, la moyenne établie par les méthodes multilatérales ne couvre qu'une section transversale. La question de savoir si le calcul de la moyenne dans le temps peut contribuer à atténuer le bruit et à amplifier la composante signal mérite d'être étudiée plus avant.

L'article d'**Isabelle Léonard et ses co-auteurs** propose une exception notable du niveau de détail des indices diffusés, en mesurant les différences entre les niveaux des prix à la consommation dans différentes régions de France métropolitaine, se concentrant tout particulièrement sur les produits alimentaires vendus en supermarché.

Pour conclure

Il est aujourd'hui possible de standardiser la mise en œuvre des données de caisse et des données moissonnées sur le Web dans plusieurs instituts de statistique. S'agissant des données de caisse, le jeu de données Dominick's Finer Foods (un jeu de données de caisse couvrant 7 années) est maintenant accessible à la Booth School of Business de l'Université de Chicago, pour développer les capacités de traitement de ce type de données[1]. Plusieurs ateliers ont été organisés pour expliquer l'utilisation de différents outils de moissonnage qui peuvent ensuite être adaptés aux besoins[2]. Pour le calcul des indices, la version bêta d'un package R est disponible, permettant d'utiliser les méthodes les plus courantes[3].

La mise à jour à venir du *Consumer Price Index Manual* de 2004 comportera un programme de recherche, qui inclura naturellement les données de caisse et le moissonnage. Toutefois, l'approche existante n'est pas remise en question du point de vue de la théorie économique, ce qui témoigne également d'une intention d'être plus concrètement applicable. Les indices dits du « coût de la vie » reconnaissent que les quantités consommées dépendent des prix. En revanche, ils ne tiennent pas compte du fait que les consommateurs peuvent acheter pour stocker un produit lorsqu'il est mis en promotion ou soldé, contredisant ainsi le postulat de base selon lequel les biens achetés pendant une période donnée sont consommés durant cette période. La substitution entre produits est éclipsée par la substitution intertemporelle. En conséquence, les résultats tirés d'une estimation statique peuvent être trompeurs.

Pour finir, les données de caisse et les données moissonnées représentent un échantillon non probabiliste, certes « grand » mais biaisé, et non la population. Il y a des transactions qui entrent dans le champ mais qui ne sont pas enregistrées électroniquement, qui ne sont

1. *https://github.com/eurostat/dff*
2. *https://unstats.un.org/bigdata/taskteams/scannerdata/workshops/Presentation_webscraping_Bogota_Statistics%20Belgium.pdf*
3. *https://cran.r-project.org/package=IndexNumR*

pas à la disposition des instituts de statistique, qui sont supprimées à l'étape du filtrage, qui ne peuvent être appariées ou liées, etc. Après tout, avoir plus de données n'est pas nécessairement mieux, c'est avoir de meilleures données qui est mieux. Les données de caisse et les données moissonnées sur le Web peuvent être extrêmement détaillées, mais leur précision peut aussi être limitée. Il serait dangereux d'accorder une confiance aveugle à ces nouvelles sources et de penser qu'elles apportent automatiquement de meilleures réponses. En réalité, les « big data » ne collectent pas la totalité mais seulement une partie des transactions, et nous ne savons pas nécessairement lesquelles sont absentes. C'est pourquoi la clé pour réduire le biais de couverture est de combiner les données classiques et les données massives. □

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

Rappel :

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes, et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

** 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*).

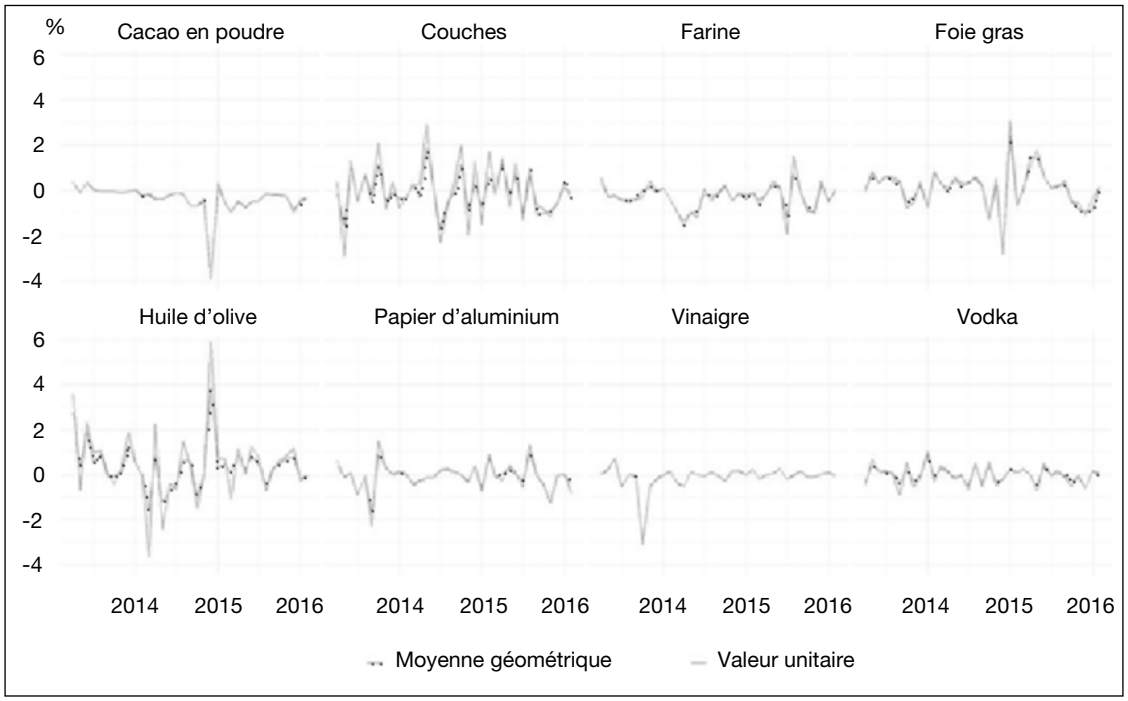
Le choix d'une formule plutôt qu'une autre a, là encore, un impact sensible sur les résultats obtenus en termes d'indice. Pour huit variétés de produits représentatives, des indices ont été construits entre 2013 et 2016 en agrégeant temporellement les prix soit grâce à une valeur unitaire ($\overline{p} = \sum_{i=1}^{28} v_{m,i} / \sum_{i=1}^{28} q_{m,i}$ 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.

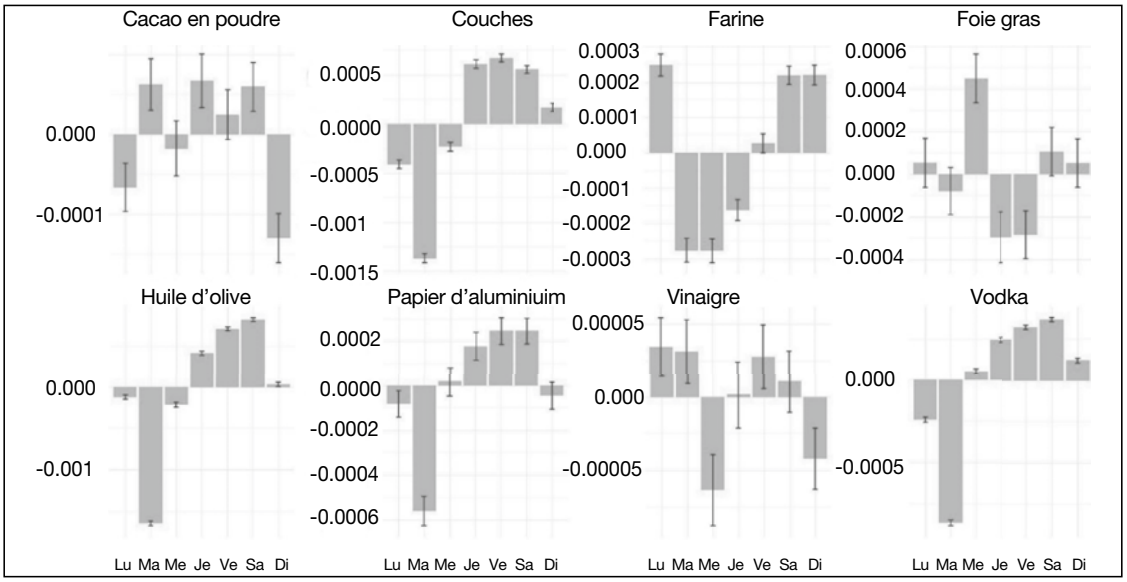
prix calculés sur une semaine montrent que les écarts de prix pour ces variétés entre jours de la semaine sont très faibles (les plus forts écarts sont de l'ordre de 0.1 %), et que pour ce type de produits, il n'y a pas eu de politique de fixation des prix différenciée au cours de la semaine sur la période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

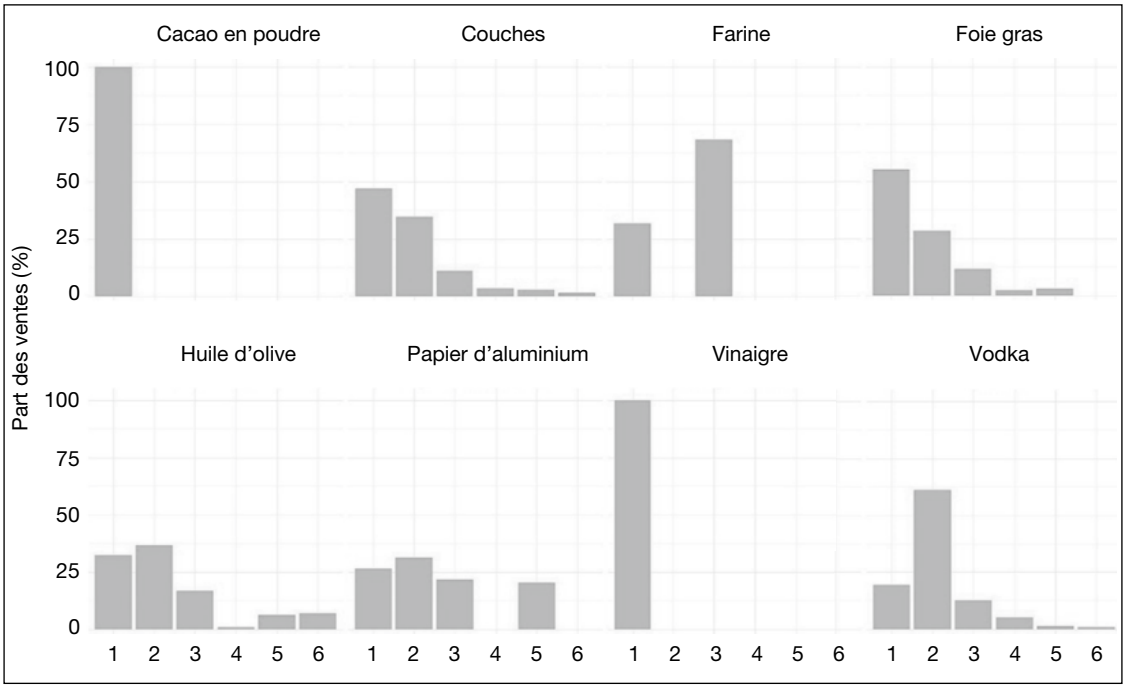
Afin de mesurer correctement les évolutions de prix, en prenant en compte ces éventuelles relances ou promotion, le panier de produits n'est pas constitué de codes-barres mais de « classes d'équivalence », des regroupements de codes-barres pour lesquels on considère que le produit est identique aux yeux 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

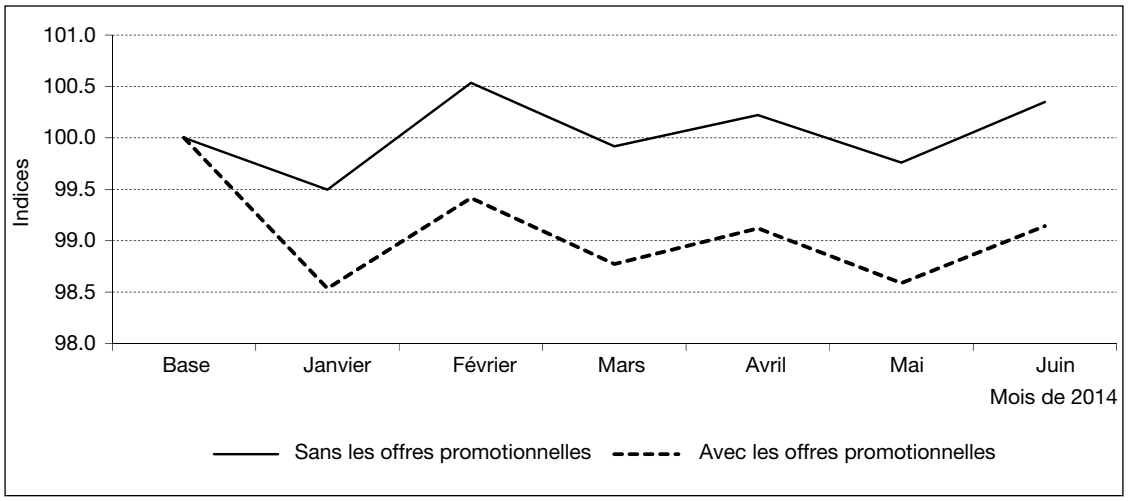
Une fois les produits identifiés par classe d'équivalence grâce à une combinaison du code-barres et du référentiel des articles, reste encore la tâche de classer les produits par variétés puis dans la nomenclature de fonction de consommation. Cette tâche est nécessaire d'une part pour des problèmes de 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

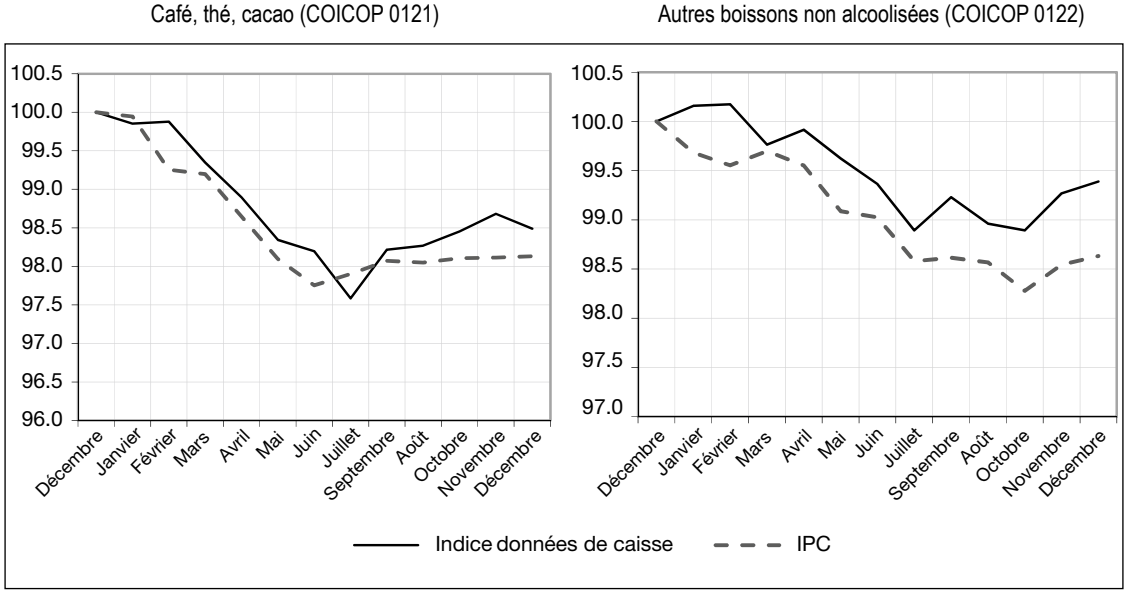


Note : en prenant en compte les offres promotionnelles, l'indice des prix des poissons fumés a chuté de 1.5 % en janvier 2014.
Champ : poissons fumés au rayon frais.
Source : données de caisse de 4 enseignes représentant 30 % du marché, en 2014.

* *
*

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

Mesure de l'inflation avec des données de caisse et un panier fixe évolutif

Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket

Can Tongur*

Résumé – Statistics Sweden a introduit les données de caisse dans certaines parties de l'indice des prix à la consommation (IPC) il y a plusieurs années, avec la préoccupation de s'assurer de la comparabilité au fil du temps et entre pays. Cet article traite de la préservation de l'approche du panier fixe et interroge la pertinence de la méthode classique du remplacement manuel d'articles, accompagné d'ajustements de la qualité et de la quantité, pour assurer la comparabilité malgré le changement de mode de collecte et l'étendue des données. Les biais de l'IPC dus à des ajustements erronés de la quantité sont analysés et illustrés avec des exemples numériques basés sur des évolutions réelles du marché suédois des produits de consommation courante. Les ajustements manuels de la qualité et de la quantité sont mis en œuvre sur un échantillon aléatoire restreint d'articles représentatifs, c'est-à-dire un panier fixe, ce qui engendre de l'imprécision ou de la variance dans l'IPC. Cette approche pourrait être mise en question compte tenu du caractère de quasi-recensement des données de caisse, et le compromis biais-variance est donc analysé dans cette étude. La variance liée à la taille de l'échantillon est estimée avec une méthode *jackknife* et comparée avec des ajustements de la qualité/quantité.

Abstract *– Statistics Sweden introduced scanner data into parts of the consumer price index several years ago, with the concern to ensure comparability over time and between countries. In this article, we discuss the issue of preserving the fixed basket approach and whether the traditional manual item replacement strategy, with quality and quantity adjustments, is still a relevant method to ensure comparability despite the change in data collection mode and extensiveness of data. Biases from improper quantity adjustments are discussed and illustrated through numeric examples based on real changes in the Swedish market of daily necessity products. Manual adjustments of quality and quantity are implemented by following a small random sample of representative items, i.e. a fixed basket, which therefore leads to imprecision or variance in the consumer price index. This may be a questionable approach given the availability of census-like scanner data, thus the bias-variance trade off is addressed. The sample size related variance is estimated through a jackknife method and contrasted with quality/quantity adjustments.*

Codes JEL / JEL Classification : E31, C15, C83, C80
Mots-clés : données de caisse, indice des prix à la consommation, IPC, panier fixe, inflation cachée, variance *jackknife*
Keywords: scanner data, consumer price index, CPI, fixed basket, hidden inflation, jackknife variance

Rappel :

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes, et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

** Statistics Sweden (Can.Tongur@scb.se)*

L'auteur remercie Anders Norberg, consultant en chef auprès de Statistics Sweden, pour sa contribution. Cet article a été largement amélioré grâce aux suggestions des rapporteurs anonymes de la revue.

Reçu le 31 juillet 2017, accepté après révisions le 4 juillet 2018
Traduit de la version originale en anglais

Pour citer cet article : Tongur, C. (2019). Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket. *Economie et Statistique / Economics and Statistics*, 509, 31–47. https://doi.org/10.24187/ecostat.2019.509.1982

Les données de caisse du commerce de détail ont été introduites dans l'indice des prix à la consommation (IPC) suédois en 2012 et concernaient initialement les produits de consommation courante. À cette époque, le Statistiska centralbyrån (Statistics Sweden) ne se posait pas de question conceptuelle sur le volume de données à utiliser. Le processus a commencé par le remplacement, un à un, des prix collectés manuellement par les données de caisse du seul détaillant à les fournir, en conservant la structure de l'échantillon pour les magasins et les articles. Avant que les données de caisse n'entrent dans la production de l'IPC, plusieurs études internes avaient été menées afin de s'assurer que cette nouvelle source de données était conforme à une attente de base, l'absence d'impact défavorable sur l'IPC.

Au fil du temps, le volume des données de caisse incluses dans l'indice et le nombre de détaillants fournissant gracieusement ces données ont augmenté. En termes de chiffre d'affaires, les données de caisse couvrent désormais plus de 80 % du marché suédois des produits de consommation courante[1]. Conséquence positive de cette expérience menée dans le secteur des produits de consommation courante, d'autres parties de l'IPC suédois sont maintenant produites à l'aide des nouvelles sources de données de transactions. Malgré la hausse des volumes de données pouvant être utilisés, surtout dans le secteur des produits de consommation courante, la production de l'IPC suédois se poursuit conformément à une stratégie bien établie d'échantillonnage des produits et des magasins. La stratégie d'échantillonnage, pour l'essentiel indépendante du mode de collecte des données, a été adaptée et l'introduction de cette nouvelle source de données alternative, extrêmement prometteuse, n'a entrainé que des changements méthodologiques mineurs – et de simples petites divergences.

À l'ère des *Big Data*, dont la popularité se répercute sur la méthodologie statistique, cette position quelque peu prudente de Statistics Sweden peut être questionnée : pourquoi ne pas utiliser en continu la totalité, ou le plus grand nombre possible, de données, ce qui semblerait intéressant et plus au goût du jour ? On s'intéresse dans cet article à la question de la préservation de la méthodologie conventionnelle de production de l'IPC en présence des données de caisse. Statistics Sweden a choisi d'utiliser ces données tout en maintenant l'approche conventionnelle, afin de traiter de façon satisfaisante le phénomène des relances commerciales. En effet, lors de ces relances, certaines caractéristiques d'un produit changent (par exemple sa taille). Le prix du nouveau produit, bien que quasiment semblable au produit antérieur à la relance, doit être ajusté en fonction du changement de quantité afin de préserver la comparabilité au fil du temps. L'impact d'une évaluation erronée des ajustements qualité/quantité sera examiné, dans le cadre de l'utilisation de paniers « automatiques » (*infra*) avec les données de caisse.

Cet article vise à étudier le compromis entre la précision de l'inflation mesurée et le biais qui survient lorsque les ajustements de la quantité sont ignorés. Bien qu'il se concentre sur les produits de consommation courante, l'analyse est également valide pour l'IPC global.

L'article s'articule comme suit. La prochaine section donne une vue d'ensemble de l'utilisation des données de caisse par Statistics Sweden dans la production de l'IPC. Il s'agit d'une section descriptive de ce mode de collecte des données relativement nouveau, et elle s'adresse principalement aux lecteurs peu familiers du sujet. Dans la section suivante, un estimateur de variance *jackknife* est appliqué dans un cadre simplifié afin d'évaluer la variance des indices de prix à la consommation avec des données de caisse. Nous examinons ensuite la question des ajustements qualité/quantité, qui est décrite et illustrée par des exemples basés sur des changements réellement survenus sur le marché suédois des produits de consommation courante. L'article se termine par des remarques générales et une mise en perspective des résultats présentés.

Les données de caisse pour les produits de consommation courante dans l'IPC suédois

Cette section présente certaines des questions méthodologiques qui ont dû être traitées avant de mettre en œuvre les données de caisse. Mais elle fait d'abord un petit détour par la terminologie et quelques éléments sur l'arrivée des données de caisse il y a une vingtaine d'années.

1. Les statistiques de marché peuvent être obtenues auprès de l'Institut de recherche commerciale suédois Handelns Utredningsinstitut *(HUI Research, 2017).*

Données de caisse, données de transactions et données massives

Dans le contexte de la vente de biens de consommation, les « données de caisse » sont une expression, peut-être un peu maladroite, désignant les « données de transactions » dans le cadre des ventes au consommateur[2]. Le terme « caisse » vient des codes-barres[3] collés sur les emballages, qui sont scannés à la caisse du magasin afin d'enregistrer les articles au point d'achat. Le terme plus général de « données de transactions » peut être utilisé de façon interchangeable dans la mesure du possible, car il a une portée plus large : il couvre les données numériques des ventes et de la consommation de biens et de services. Les données de transactions relatives aux ventes sont, pour la plupart, des données bien structurées découlant d'un système commercial, et ne doivent pas être confondues avec, par exemple, les « données en masse » non structurées. Les données de transactions peuvent être de grande envergure, à haute fréquence (et elles peuvent être obtenues quasiment en temps réel), et ont un caractère proche des données administratives, en ce sens qu'elles ne sont pas destinées à établir de statistiques publiques, mais plutôt destinées à la gestion des stocks ou au suivi des ventes ou des bénéfices.

L'arrivée des données de caisse dans l'IPC suédois

Les données de caisse ne sont pas totalement nouvelles pour Statistics Sweden. Au milieu des années 1990, alors que les données numériques faisaient leur apparition, des contacts avaient été pris avec des analystes du marché suédois afin d'avoir un premier aperçu sur cette nouvelle source de données prometteuses – dont les avantages potentiels pour l'IPC étaient évidents et attrayants. Néanmoins, ces données étaient très chères et donc hors de portée pour une agence gouvernementale dans le contexte de la crise économique nationale la plus sévère depuis la guerre (pour plus de détails économiques et politiques, voir par exemple Bäckström, 1997, ou Englund, 2015). Aujourd'hui, une bonne vingtaine d'années plus tard, cette source de données fait partie intégrante et constitue une étape naturelle de la collecte de données mensuelle effectuée pour l'IPC suédois, et Statistics Sweden reçoit à titre gracieux les données de nombreux détaillants sur la base d'accords bilatéraux à but non lucratif. Cela ne concerne que l'échantillon de magasins inclus dans l'IPC durant une année donnée. Dans la mesure où les enseignes fournissent leurs données gratuitement, Statistics Sweden a limité ses demandes, ce qui est par ailleurs un facteur de confiance car les détaillants ne fournissent pas d'informations commerciales exhaustives à haute fréquence.

Panier de l'IPC, données relatives aux transactions et exceptions

Le panier de l'IPC

Le panier de l'IPC est présenté au tableau 1 selon la nomenclature internationale COICOP[4] à deux chiffres. Les prix sont collectés pour des produits spécifiques au sein de ces catégories de consommation. Plusieurs étapes de calcul interviennent entre la valeur totale de l'IPC et les produits définis – l'IPC est tout simplement une hiérarchie dans laquelle les données sur les prix sont agrégées par étape.

Un produit donné chez un détaillant donné, qui fait l'objet de la mesure du prix, est dénommé « offre de produit ». Les prix observés sont regroupés grâce à des formules d'indice et selon des poids infra-annuels fixes pour les groupes de produits, qui sont souvent des indices de premier niveau, c'est-à-dire des agrégats élémentaires. Un exemple de groupe de produits est le lait : les prix des différentes variétés de toutes les marques, tous les magasins et tous les types de teneur en matière grasse sont regroupés dans un même groupe de produits. Il en va de même pour les boissons gazeuses, avec ou sans sucre et toutes tailles confondues.

Le poids d'un groupe de produits reflète la part que ce groupe occupait dans la consommation privée l'année complète précédant l'année de base de l'indice, considérée comme période de référence. L'année de base de l'indice est décembre de l'année y–1 et les mois en cours utilisés pour relever les prix correspondent à l'année y, de sorte que les poids s'entendent (normalement) à partir de l'année y–2 pour l'indice mensuel. L'IPC est une série d'indices chaînés s'étalant sur plusieurs années, et une partie de la discussion présentée dans cet article concerne le chaînage mensuel (infra-annuel).

---

2. Le Manuel de l'IPC fait la différence entre les données de caisse et les données recueillies au point de vente électronique (§6.117, ILO 2004), ce que cet article ne fait pas.

3. Le code-barres relie l'article, par le biais de son emballage, à un numéro d'article distinct fourni par un acteur du marché international conformément à la norme EAN (European Article Number) */GTIN (*Global Trade Item Number*).*

4. Nomenclature COICOP (Classification des fonctions de consommation des ménages). Voir la page du site Web des Nations Unies dédiée à ce sujet (UN, 2017).

Données de transactions au sein du panier

Les données de transactions sont utilisées pour mesurer les prix de certaines catégories de produits et sont également une source d'information pour le calcul des poids. Dans le cadre des produits de consommation courante, elles comprennent le chiffre d'affaires hebdomadaire au niveau de l'article et du magasin, il s'agit donc d'informations sur la consommation réelle. Certains produits, comme par exemple les boissons alcoolisées, les médicaments vendus en pharmacie et les soins dentaires, ne sont couverts qu'à une fréquence mensuelle par des données exhaustives. En outre, Statistics Sweden dispose des données de caisse agrégées par année pour l'ensemble de la Suède depuis le milieu des années 1990 ; elles ont servi à construire le panier.

Comme on peut le voir au tableau 1, si les données de transactions sont utilisées, elles ne le sont pas dans toutes les parties du panier (les principales exceptions sont indiquées à l'encadré 1).

Mise en œuvre des données de caisse dans l'IPC suédois

Dans la mesure où Statistics Sweden traite les données de caisse depuis plus de cinq ans pour produire l'IPC mensuel, nous proposons de commencer par expliquer certains des choix faits par le passé.

Options d'utilisation des données de caisse

Statistics Sweden et le Conseil de l'IPC ont décidé en 2011 de conserver l'approche du panier fixe

Tableau 1
Poids dans le panier de l'IPC pour l'année 2016

Code	Titre	Poids dans le panier (‰)	Données relatives aux transactions
01	Produits alimentaires et boissons non alcoolisées	139	Oui
02	Boissons alcoolisées, tabac et stupéfiants	39	Oui
03	Articles d'habillement et chaussures	53	Non
04	Logement, eau, gaz, électricité et autres combustibles	251	Oui
05	Meubles, articles de ménage et entretien courant du foyer	55	Non
06	Santé	38	Oui
07	Transports	135	Non
08	Communications	35	Non
09	Loisirs et culture	120	Non
10	Enseignement	5	Non
11	Restaurants et hôtels	67	Non
12	Biens et services divers	63	Oui
Total	IPC	1 000	

Note : conformément aux divisions de la nomenclature COICOP (deux chiffres) pour la consommation des ménages. Les données de transactions sont indiquées dès lors qu'elles sont utilisées dans les relevés de prix. Deux autres divisions COICOP (codes 13 et 14) couvrent la consommation autre que celle des ménages et sont donc hors du champ d'application de l'IPC.

Encadré 1 – **Exceptions dans les données de caisse des produits de consommation courante : détaillants ne fournissant pas de données et produits frais**

Dans les deux premières divisions de la nomenclature COICOP (01 et 02), les données de transactions sont utilisées de façon quasi exclusive, à deux exceptions près. Premièrement, certains détaillants inclus dans la division 01 (Produits alimentaires et boissons non alcoolisées) ne fournissent pas de données de transactions ce qui nécessite donc une collecte manuelle des prix. Deuxièmement, la collecte manuelle des prix se poursuit dans le segment des fruits et légumes frais, de la viande fraîche et des fromages. Ces articles sont habituellement vendus au poids, ou parfois à la pièce (par exemple, unités d'avocats ou de citrons).

En 2017, les données de caisse ont fait leur apparition pour les produits frais, d'abord pour un seul détaillant (Tongur & Sandén, 2016). En 2018, la double collecte des données (manuelle et numérique) a pris fin et les données de caisse ont été entièrement adoptées pour les détaillants fournissant des données de caisse (Bilius *et al.*, 2017). Des sujets connexes sont traités dans les publications de l'office des statistiques Norvégien (Statistisk sentralbyrå), voir Nygaard (2010) ou Rodriguez & Haraldsen (2005), ainsi que dans celles de CBS (Pays-Bas), voir van der Grient & de Haan (2010).

(voir encadré 2), estimant que c'était la façon la moins intrusive d'inclure les données de caisse[5]. Sa mise en œuvre a été immédiate, dès 2012, et a plus ou moins consisté à simplement changer le mode de collecte des données, ce qui était censé minimiser l'impact sur la production de l'IPC global et les systèmes informatiques associés. Cette décision a été prise sur la base de plusieurs études et analyses de données, ainsi que sur des comparaisons avec la collecte de prix manuelle (Norberg *et al*., 2011). À côté de la question de savoir comment utiliser les données de caisse dans la pratique, il fallait également décider si les données devaient effectivement être utilisées. Norberg *et al*. (2011) ont identifié quatre façons fondamentalement différentes d'utiliser les données de caisse, en se limitant au champ des produits de consommation courante. Elles sont présentées dans l'encadré 3.

5. Cette décision a été prise avec l'approbation du Conseil de l'IPC, qui avait un mandat réglementaire à ce moment.

Encadré 2 – **Le Conseil de l'IPC suédois**

Le Conseil de l'IPC suédois (*Nämnden för Konsumentprisindex* en suédois) est un organe consultatif scientifique et interdisciplinaire externe dédié aux méthodologies de production de l'IPC. L'IPC suédois est une statistique produite chaque mois qui n'est pas révisable. Le Conseil se réunit habituellement deux fois par an dans les locaux de l'institut (Statistiska centralbyrån – Statistics Sweden).

Le Conseil a été créé il y a plusieurs dizaines d'années et joue depuis 2017 un rôle de consultation, sans force d'obligation, pour les questions ayant une importance substantielle dans le cadre de l'IPC. Ses membres sont nommés par Statistics Sweden et représentent les établissements publics concernés par l'IPC, comme la Riksbanken (banque centrale de Suède) et d'autres agences gouvernementales ou organismes publics. L'IPC norvégien y est aussi représenté, afin de partager les expériences et de renforcer la collaboration entre les pays nordiques. Ces collaborations ont été utiles lors de l'introduction des données de caisse, puisque l'institut norvégien (Statistisk sentralbyrå) fait partie des premiers instituts nationaux à les avoir utilisées. Des experts externes de portée internationale sont également membres du Conseil.

Avant 2017, le mandat du Conseil avait force d'obligation. Par ses décisions, le Conseil pouvait influer sur les réponses apportées à toute question relative à l'IPC. En outre, ses décisions ne pouvaient pas être officiellement remises en question, conformément aux instructions légales applicables à Statistics Sweden. Parallèlement, le Conseil comptait un membre permanent du ministère de tutelle. Toutefois, en 2012, le système de la statistique officielle en Suède, ainsi que le rôle de Statistics Sweden en tant que grande agence gouvernementale ont été réexaminés (SOU, 2012). Cet examen avait été demandé par le gouvernement et, s'agissant de l'IPC, a été recommandé de supprimer la force obligatoire du mandat du Conseil de l'IPC, qui semblait discutable en termes d'objectivité de l'agence et contraire au Code des bonnes pratiques de la statistique européenne.

Encadré 3 – **Quatre façons d'utiliser les données de caisse dans la production de l'IPC**

A – Remplacement des données de prix collectées manuellement par des données de caisse pour les échantillons ordinaires de magasins et de produits : cela n'engendrerait que des changements/ajustements mineurs dans le mode actuel de production de l'IPC, et la réglementation relative aux indices des prix à la consommation harmonisés[(a)] serait entièrement respectée.

B – Utilisation des données de caisse en tant qu'information accessoire : cela exigerait de choisir entre deux approches possibles et de continuer à échantillonner manuellement les prix cotés. L'échantillon serait calibré selon (i) les données de caisse des périodes correspondantes, ou (ii) les données de caisse seraient calibrées selon la collecte manuelle correspondante.

C – Calcul de l'indice à partir d'un recensement de tous les produits pour lesquels des données de caisse sont disponibles : soit l'approche du panier fixe est appliquée à grande échelle, avec l'attrition correspondante du panier durant l'année, soit une méthodologie totalement différente est introduite, probablement en adaptant les méthodes néerlandaise ou norvégienne[(b)] avec un chaînage mensuel.

D – Utilisation des données de caisse pour l'audit et le contrôle qualité : il s'agit de l'usage le plus minimaliste des données de caisse dans la production de l'IPC. Toutefois, si elles n'étaient utilisées que pour cet usage, cela représenterait également un énorme gaspillage de ressources.

(a) Voir la réglementation sur les indices des prix à la consommation harmonisés (Eurostat, 2013).
(b) Comme le soulignent van der Grient & de Haan (2010), Nygaard (2010) et les premières discussions avec le Statistisk sentralbyrå (institut de la statistique de Norvège).

Les quatre options présentées à l'encadré 3 donnent de possibles utilisations de ces données, pour des usages allant au-delà du simple contrôle qualité des prix collectés manuellement qui correspond à l'option D. L'option B a été envisagée mais elle est apparue peu optimale par rapport aux autres. Les données de caisse ayant été obtenues et mises en œuvre de façon progressive, la première option (option A) maintenait le statu quo pour la production de l'IPC s'agissant du calcul de l'indice et de la conception des échantillons. L'opportunité présentée par les données massives et l'émergence de nouvelles méthodes dans ce domaine, dont CBS (Bureau central de la statistique aux Pays-Bas) et Statistics Norway étaient les pionniers rendaient l'option C attractive. Toutefois, face à des contraintes temporelles et économiques et compte tenu du besoin d'acquérir de l'expérience avec cette nouvelle source de données, l'option A semblait préférable pour commencer la transition vers ces nouvelles données. Utiliser l'ensemble des données (option C) n'était pas la solution préférable pour la première étape, mais elle demeure le but.

Panier fixe / panier dynamique

Les données de caisse, en 2012, ont été pour la première fois mises en œuvre dans le cadre d'un panier fixe standard. D'autres pays appliquent une approche moins conventionnelle, à savoir le panier dynamique. Les principales caractéristiques des deux approches sont présentées dans les consignes d'Eurostat pour le traitement des données de caisse de supermarchés (Eurostat, 2017a). Ces consignes ont été formulées à partir des informations fournies par les pays participants, afin de formaliser les approches appliquées par les différents pays et de s'efforcer d'harmoniser les indices des prix à la consommation dans les pays commençant à utiliser des données de caisse. Les principales différences, et les principaux avantages et inconvénients des deux approches sont présentés ci-dessous.

L'approche du panier fixe

L'approche du panier fixe signifie que, chaque mois t (ou trimestre) d'une année donnée y, le panier est maintenu dans une configuration aussi constante que possible. Les prix des articles inclus dans le panier sont observés (si cela est possible) puis comparés de façon directe chaque mois avec les prix concernés du mois de base des mesures annuelles, qui correspond habituellement au mois de décembre y–1.

L'évolution constante du panier et le problème des remplacements

L'inconvénient de cette approche prudente est de ne pas tirer parti de l'abondance des données, ni de la mise à jour des informations de marché. Elle repose sur un panier qui pour être maintenu dans la même configuration, est limité – cette contrainte venant de sa révision mensuelle, c'est-à-dire du remplacement de produits disparus. Le problème des remplacements est central pour la préservation de la comparabilité dans le temps, et constitue sans doute l'argument le plus convaincant pour maintenir l'approche classique : les changements de qualité et de quantité intervenant lors des remplacements sont traités de façon explicite. Lorsque les données relatives à certains articles ne sont pas observables, il faut faire un choix entre le remplacement par un autre article comparable (ce qui, dans le meilleur des cas, passe par une relance du même article) ou, si cela n'est pas possible, l'arrêt de l'article. Dans les cas extrêmes, l'attrition peut engendrer un panier non représentatif[6] basé sur les articles restants. Le problème peut être contourné (mais pas résolu) grâce à une solution plus automatisée pour les données de caisse : le panier dynamique.

L'approche du panier dynamique

Dans le cadre de l'utilisation dynamique des données de caisse, les prix sont mesurés pour un panier mis à jour en continu. Pour cela, un indice mensuel des articles appariés est calculé pour les ratios de prix des articles exactement appariés entre des mois adjacents (t et t–1 de l'année y), et ce chaînage mensuel est ensuite relié au mois de base de l'indice (décembre de l'année y–1). Cette approche revient à un panier fixe si tous les articles (et les poids) sont identiques durant toutes les périodes – voir par exemple Eurostat (2017b, formules 8.11 et 8.14), Eurostat (2017a) et Fisher (1922).

L'approche dynamique retient dans le panier l'univers d'articles le plus récent, c'est-à-dire qu'il s'agit d'un échantillon à jour, et cette couverture est la plus représentative et complète qui soit. Comme Boskin *et al*. (1997), entre autres, le soulignent, une telle approche devrait être utilisée pour réduire les coûts de collecte et pour élargir l'assortiment des biens et services inclus dans l'IPC.

6. Dans ce cas, le panier ne couvre pas toute la consommation cible et n'est donc pas représentatif.

À des fins de régularité, c'est-à-dire en termes de 1) stabilité, 2) représentativité au fil du temps et 3) parcimonie des données visant à éviter le bruit, il est nécessaire d'exclure du panier les produits dont la part dans la consommation au cours du mois concerné est trop faible, comme expliqué par Eurostat (2017a) et par van der Grient & de Haan (2010), ou des produits en fin de cycle qui peuvent être bradés. Même avec ces précautions, un biais de chaînage peut découler des variations de prix, hausses ou baisses significatives durant certaines périodes, tirant ainsi l'indice vers le haut/bas durant la période concernée. Quand les prix des produits reviennent à leur niveau initial sans que l'indice ne retrouve son niveau précédent, il y a une dérive de l'indice.

Le problème peut être illustré comme suit. Supposons, par exemple, qu'un filtre de taille soit appliqué de sorte que les dix premiers articles en termes de chiffre d'affaires soient choisis pour un mois donné (lesquels étaient déjà retenus dans le panier la période précédente). Si le chiffre d'affaires d'un article est élevé un mois donné, cela peut être simplement parce qu'il est régulièrement très consommé, mais cela peut aussi être « temporaire » en raison, par exemple, d'une campagne publicitaire ou de ventes saisonnières (par exemple à Noël). Le mois suivant, ces articles « temporaires » sont susceptibles de ne plus être vendus au même prix, ou d'être arrêtés. En conséquence, les mêmes articles ne seront pas inclus dans les dix premiers ou le seront à des niveaux de prix différents, et l'indice chaîné ne retrouvera pas son niveau précédent, c'est-à-dire qu'il subira une dérive.

Cette dérive est encore plus marquée quand les quantités vendues, connues grâce aux données de caisse, sont utilisées dans la formule d'indice pour agréger les prix. Le biais de chaînage, une question à part entière, a été examiné de façon approfondie (voir Johansen & Nygaard, 2011 ; Nygaard, 2010 ; van der Grient & de Haan, 2011).

L'approche dynamique et les remplacements/ relances : un problème qui n'en est pas un

Le plus gros inconvénient de l'approche dynamique est qu'elle ne retient pour le calcul de l'indice d'un mois donné que les produits présents deux mois successifs. Or une relance commerciale s'accompagne en général d'une augmentation de prix (soit le prix est inchangé pour une quantité vendue plus faible, soit le prix augmente sans amélioration tangible de la qualité). Cette augmentation est ignorée, ou « masquée », si la relance n'est pas traitée explicitement. En effet, avec l'approche dynamique, aucun ajustement qualité n'est pratiqué car tous les articles du panier dynamique sont par définition présents deux mois adjacents. Or, comme le note Eurostat, « *les relances et les remplacements sont potentiellement problématiques dans cette méthode, car le système ne relie pas automatiquement un code d'article sortant avec le code de la relance ou du remplacement.* » (Eurostat, 2017a, p. 28).

Les données hebdomadaires dans un indice mensuel : comment les agréger ?

Disposer de données à plus haute fréquence pose la question des données multiples : faut-il combiner des données collectées à des dates différentes et si oui, comment ? La collecte manuelle des prix était effectuée une fois par mois et par magasin, ce qui est toujours le cas. Les prix collectés sont donc des prix ponctuels uniques. Comme stipulé dans les consignes sur les indices des prix à la consommation harmonisés (Eurostat, 2013), la procédure opérationnelle standard consiste à relever les prix durant la semaine correspondant au 15 du mois, ou également une semaine avant/après le milieu de semaine. Les relevés de prix (c'est-à-dire ceux des magasins inclus dans l'échantillon) sont habituellement attribués *a priori* durant trois semaines, pour une plus grande précision sur le mois.

Les données de caisse offrent la possibilité de calculer la consommation hebdomadaire, c'est-à-dire le chiffre d'affaires et les quantités achetées chaque semaine. Les données suivent les semaines civiles (du lundi au dimanche), ce qui empêche une utilisation régulière sur plus de trois semaines en raison des semaines chevauchant deux mois. En utilisant le milieu de semaine et les deux semaines adjacentes, on obtient trois points de données par offre de produit, dans le meilleur des cas. L'échantillon est donc plus précis, mais dans une dimension peu utilisée dans la méthodologie standard en raison de la nature des statistiques économiques : mesure discrète de données temporelles continues (voir le Manuel de l'IPC § 15.70, ILO, 2004).

Deux possibilités intuitives pour combiner les points de données hebdomadaires en un seul prix unique par offre de produit et par mois sont la moyenne géométrique et la moyenne arithmétique, qui semblent toutes les deux appropriées. Lors de la toute première mise en œuvre, le Conseil de l'IPC a décidé qu'une moyenne géométrique non pondérée

sur la période (maximale) de trois semaines serait appropriée pour calculer le prix mensuel de chaque offre de produit à partir des données de caisse. Ainsi, les données de caisse fournies par un même détaillant correspondraient aux données autres que les données de caisse des autres offres de produit. L'idée était que les trois semaines de données de caisse pouvaient être considérées comme trois cycles de collecte de données plutôt que comme une collecte unique, comme les autres offres de produit. L'agrégation par moyenne géométrique non pondérée était également conforme à la construction de l'indice, qui passe par une moyenne géométrique (indice de Jevons).

La question de l'agrégation semaine/mois a été réexaminée après obtention de données d'un plus grand nombre d'enseignes, à partir de 2013, et le Conseil de l'IPC a été consulté (Sammar & Norberg, 2012). Cette fois-ci, en raison de l'élargissement de la couverture, le Conseil a décidé d'utiliser une moyenne arithmétique pondérée sur trois semaines, qui refléterait mieux les prix unitaires mensuels, en ligne avec les données réelles (hebdomadaires). « Pondéré » signifie que les chiffres d'affaires de trois semaines (au plus) sont additionnés puis divisés par la somme des quantités de ces semaines, donnant ainsi un prix unitaire mensuel moyen.

L'évolution des deux moyennes envisagées a été étudiée dans le contexte d'un indice de prix (cf. Norberg *et al*., 2012), ce qui a permis d'identifier des comportements différents dans certaines situations. Pour plus de 90 % des observations, les deux moyennes ne différaient que très peu, mais les différences s'accentuaient lorsque la pondération jouait un rôle important, par exemple en période de vacances lorsque les prix étaient bas. L'étude a également permis de réaliser que tout impact sur la période de base affectait par la suite les rapports de prix agrégés (c'est-à-dire l'indice) tout au long de l'année, même si les deux moyennes coïncidaient durant le mois concerné.

Maintenance des échantillons

Suite au passage aux données de caisse, les remplacements/substitutions d'articles correspondant aux articles du panier devenus obsolètes devaient être faits en interne par les responsables de l'IPC, à partir de la surveillance de l'attrition au sein du panier. Afin de limiter l'épuisement potentiel des échantillons, un système de maintenance du panier extrêmement simple a été mis en œuvre, permettant de comparer les ventes du mois en cours t avec la période de base décembre y–1. Ce suivi se fait tant au niveau du nombre de magasins dans lesquels le produit a été vendu qu'au niveau du nombre de lots vendus, c'est-à-dire qu'il s'agit d'une analyse bidimensionnelle. Cela se déroule *a posteriori* pour chaque mois achevé. Ainsi, l'échantillon utilisé pour l'IPC reste (vraisemblablement) représentatif, ce qui ne demande qu'une journée de travail chaque mois, au plus, pour rechercher les articles substitués dans les données de caisse. Les prix manquants ne font l'objet d'aucun calcul et les magasins ne sont pas remplacés s'ils ferment avant la prochaine date de mise à jour de l'échantillon. Toutefois, les cas de non-réponse sont rares, surtout s'il s'agit de magasins bien établis ou dont le chiffre d'affaires est élevé.

Estimation de la variance de l'indice

On s'intéresse ici à la contribution d'un article à la variance de l'indice des prix dans le cas d'un panier fixe, en utilisant les données de caisse dans leur intégralité ou partiellement. Après une brève description de l'échantillonnage, la construction de l'indice dans le cadre des agrégats élémentaires est présentée, suivie par une estimation *jackknife* de la variance. Cette section se termine par une discussion sur les propriétés de population finie de l'échantillon des produits de consommation courante.

Échantillonnage des articles et des magasins pour les produits de consommation courante

La conception de l'échantillon porte sur deux dimensions : l'endroit et le produit (articles pouvant être achetés). Par « endroit » il faut entendre le magasin dans lequel les produits destinés à une consommation privée sont achetés. Les articles sont sélectionnés par le biais d'un échantillonnage annuel, quel que soit le mode de collecte des données. Pour les données de caisse comme pour la collecte manuelle des prix, une probabilité d'échantillonnage proportionnelle à la taille est appliquée dans les deux dimensions (voir Ohlsson, 1990 ; Rosén, 2000).

Échantillonnage des articles

À partir de chacune des enseignes couvertes par des données de caisse, environ 800 articles sont inclus dans l'échantillon annuel. Les bases de sondage sont définies chaque année en fonction des données de caisse annuelles agrégées de l'année précédant le mois de base. Les codes identifiant les articles dans les données de caisse, les codes EAN/GTIN et les niveaux plus détaillés

de la nomenclature COICOP sont ensuite reliés de façon précise. En faisant les appariements nécessaires avec les données de caisse hebdomadaires, on obtient l'échantillon souhaité. Les échantillons d'articles des enseignes sont tirés avec une coordination négative entre les échantillons des différentes chaînes. Toutefois, de nombreux articles de marques répandues sont vendus par tous les détaillants et représentent de gros volumes de vente. Ces articles sont souvent inclus dans plusieurs des échantillons spécifiques aux détaillants.

Échantillonnage des magasins

L'échantillon des magasins de produits de consommation courante rassemble environ 60 magasins, répartis dans l'ensemble du pays. La probabilité d'inclusion dans l'échantillon est proportionnelle à la taille (*Poisson sampling*, voir Ohlsson, 1990). Ainsi, il est possible d'obtenir des rotations des points de vente tirés. Toutefois, le système de rotation standard de Statistics Sweden (20 % chaque année) n'est pas strictement appliqué ici. La rotation est appliquée si elle se justifie d'un point de vue probabiliste (c'est-à-dire en termes de représentativité), afin d'éviter une trop forte pression sur les fournisseurs de données pour qu'ils changent le contenu des données communiquées. Pour des raisons techniques, les magasins sont rééchantillonnés chaque année, mais ne sont remplacés que si leur importance relative est extrêmement affectée par rapport à l'échantillonnage des années précédentes.

Configuration de l'estimation

Il est très difficile d'estimer la variance dans un indice des prix à la consommation. La variance découle de l'échantillonnage bidimensionnel, magasins et articles ; des évaluations formelles de la variance sont fournies par Balk (1989, 1991), Dalén & Ohlsson (1995) et Norberg (2004).

L'indice au niveau le plus détaillé : agrégats élémentaires

Les agrégats élémentaires sont calculés comme la moyenne géométrique[7] des rapports de prix d'articles appartenant à un groupe de produits donné, dans tous les magasins. Les rapports de prix de la période d'observation t de l'année en cours y par rapport aux prix du mois de base 0, $P_{t,i}$ et $P_{0,i}$ constituent la formule de l'indice $I_g^{0,t}$:

$$I_g^{0,t} = \prod_{i=1}^{k_g} \left(\frac{P_{t,i}}{P_{0,i}} \right)^{w_i} \quad (1)$$

où la somme est calculée pour les k_g offres de produit i d'un groupe de produits g dans lequel chaque offre de produit peut avoir un poids distinct w_i. Dans le cas suédois, les poids w_i sont calculés en fonction des probabilités relatives aux magasins et aux articles. Il s'agit pour la plupart de poids unitaires, c'est-à-dire égaux ($w_i = 1$), mais certains peuvent être plus importants pour refléter un article largement vendu dans un hypermarché, par exemple une marque de café.

Si tous les poids sont égaux (ce qui revient à ne pas pondérer), alors l'équation (1) correspond à l'indice de Jevons non pondéré. Si les éléments inclus dans l'échantillon résultent d'un échantillonnage proportionnel à la taille, les probabilités d'inclusion et les poids se compensent mutuellement, c'est-à-dire que l'on obtient une pondération implicite. Lorsque les poids reflètent la part de la consommation respective des articles, l'expression correspond à l'indice de Young géométrique (voir le Manuel de l'IPC, formule 1.9, ILO, 2004).

La méthode du jackknife *pour l'échantillonnage stratifié*

La méthode du *jackknife* suggérée ici permet une approximation de la contribution à la variance du $n^{\text{ème}}$ élément d'un échantillon existant. La méthode est expliquée dans Wolter (1985), et une analyse similaire sur données de caisse est disponible dans l'étude de Leaver & Larson (2001) pour l'IPC américain (Bureau of Labor Statistics).

La stratégie de calcul consiste à estimer le paramètre cible, ici une expression de l'indice agrégé pour les indices du groupe de produits (équation 1) en excluant, un par un, chaque élément de l'échantillon existant, c'est-à-dire en conservant n–1 éléments à chaque estimation et en calculant le paramètre cible en fonction des éléments restants. En appliquant cette procédure à tous les éléments n, on obtient la contribution moyenne à la variance. L'échantillon de magasins choisi est fixe, c'est-à-dire que l'échantillon d'articles dépend de l'échantillon existant de magasins. Cette approche est supposée suffire pour la validation de principe, à savoir le compromis entre la

7. Cette formule d'indice est l'une des deux méthodes explicitement recommandées pour les indices des prix à la consommation harmonisés (Eurostat, 2013) au plus bas niveau.

contribution à la variance d'un article et le biais qui survient lorsque les ajustements de la quantité sont négligés.

Mécanisme d'estimation jackknife

Les quelque 800 articles échantillonnés pour lesquels des données de caisse sont disponibles dans chacune des trois enseignes constituent ensemble environ 90 groupes de produits du secteur des produits de consommation courante dans la nomenclature COICOP. Les groupes de produits sont, par définition, les agrégats élémentaires pour lesquels un indice de prix est calculé avec l'équation (1) pour tous les produits et les enseignes, c'est-à-dire qu'il y a un agrégat pour tous les articles d'un groupe de produits. Les articles sont classés et codés selon le groupe de produits auquel ils appartiennent, et donc un « article » correspond à un produit.

Le dispositif de stratification est présenté au tableau 2, et montre le mécanisme d'exclusion pour chaque cycle n–1. Dans ce dispositif, les groupes de produits sont croisés avec chaque enseigne pour définir les strates, ce qui donne environ 270 strates desquelles les articles sont exclus. L'équation (1) estimée pour tous les groupes de produits débouche sur le paramètre cible : l'indice agrégé des produits de consommation courante pour la division 01 de la nomenclature COICOP.

De par leur conception, les 90 groupes de produits croisés avec les trois enseignes au plus donnent environ L = 270 strates. Au total, les près de 800 produits échantillonnés au sein de chaque enseigne donnent 2 400 produits, avec des variations dues à l'évolution des assortiments. Une strate de magasins h compte n_h articles/produits. Le n_h varie selon les strates au sein du même groupe de produits, qui compte donc k_g produits au total : $k_g = \sum_{h=1}^{H} n_h, h \in g$. Au sein de chaque k_g il peut y avoir H = 3 strates, où h se résume à L = 270 pour tous les g : $h \in (g, L)$.

Les quelques strates où un seul produit est identifié sont éliminées du calcul car la procédure n–1 donne un nombre nul de produits restants, ce qui veut dire qu'aucune variance ne peut être estimée dans la strate concernée. Les assortiments et les échantillons variant d'une enseigne à l'autre, parfois de manière substantielle, tous les groupes de produits ne sont pas nécessairement inclus dans les trois enseignes.

Chaque estimation exclut séquentiellement l'une des rangées indiquées au tableau 2, c'est-à-dire chaque produit d'une strate, de sorte qu'aucun élément aléatoire n'entre dans la procédure d'estimation. Par contre, le caractère aléatoire de l'échantillon initial est reflété entre les cycles par la modification de la composition de l'échantillon concerné.

Le paramètre d'intérêt

L'équation (1) peut être exprimée sous forme logarithmique, produisant la somme suivante pour chaque groupe de produits, suivie d'une exponentiation :

$$I_g^{0,t} = \prod_{i=1}^{k_g} \left(\frac{P_{t,i}}{P_{0,i}} \right)^{w_i} = \exp\left[\sum_{i=1}^{k_g} w_i \left(\ln\left(P_{t,i}\right) - ln(P_{0,i}) \right) \right] \quad (2)$$

Tableau 2
Configuration du mécanisme d'estimation *jackknife*

Cycle d'estimation	Groupe de produits	Code du produit	Strate *h*	Chaîne
1	1113	1113001	1	1
2		1113002	1	1
3		1113003	2	2
4		1113004	3	3
5		1113005	3	3
6		1113006	3	3
7	1114	1114001	4	1
.	.	.	.	.
.	.	.	.	.
n = 2 400	.	.	*L* = 270	.

Note : les nombres n = 2 400 et L = 270 sont approximatifs et sont fournis uniquement à titre d'illustration. Les nombres exacts sont indiqués dans la sous-section consacrée aux estimations. Les champs grisés illustrent la stratification en termes de chaîne.

La partie entre crochets à droite de l'équation (2) est une version linéarisée de l'équation (1), semblable à la formule utilisée par Leaver & Larson (2001). Cela constituera le paramètre d'intérêt lors de l'élimination des produits/articles, n–1, au sein de chaque strate h du groupe de produits g.

Pour les estimations de cette étude, l'indice de l'agrégat élémentaire (2) est calculé de façon légèrement différente de la pondération réelle[8]. La différence tient au fait que les observations et les rapports de prix au sein de chaque enseigne (=strate) sont ramenés à la moyenne et agrégés pour le groupe de produits en pondérant selon la part de marché moyenne de chaque détaillant pour obtenir (2) pour le groupe de produits complet. Cela remplace les poids des articles individuels w_i ; il est nécessaire de le faire car l'alternance du nombre de produits compense la pondération implicite découlant de l'échantillonnage proportionnel à la taille. Les poids sont normalisés, de sorte que, selon le nombre d'enseignes pour chaque groupe de produits, un poids connu *a priori* est attribué au rapport de prix moyen des détaillants[9]. Cela modifie l'équation (2) de la façon suivante :

$$I_g^{0,t} = \prod_{h=1}^{H}\left[\prod_{i=1}^{n_{h,g}}\left(\frac{P_{t,i}}{P_{0,i}}\right)\right]^{w_h} = exp\left[\sum_{h=1}^{H} w_h \sum_{i=1}^{n_{h,g}}\left(\ln\left(P_{t,i}\right) - ln(P_{0,i})\right)\right] \quad (2')$$

La dernière estimation de l'indice de prix des produits de consommation courante est une moyenne arithmétique pondérée calculée sur l'ensemble des indices des groupes de produits, selon la formule

$$I^{0,t} = \sum_{g=1}^{G} w_g I_g^{0,t} \quad (3)$$

où les poids du groupe de produits w_g sont normalisés de façon à sommer à 1 (leur part agrégée dans le panier total, cf. tableau 1).

Par analogie aux définitions de Wolter (1985) pour l'estimation par stratification, l'indice de prix de l'équation (3) est calculé lorsque la $(h,i)^{\text{ème}}$ observation est supprimée. Cela est fait pour toutes les suppressions au sein d'une strate et dans toutes les strates, soit environ n = 2 400 fois, ce qui produit autant d'estimations qu'il y a d'articles/de produits. L'estimation de la variance est obtenue avec, au plus, environ L = 270 moyennes (strates) tirées des estimations (voir équation (5) ci-dessous). Ces L moyennes sont calculées, pour chaque strate h, comme l'estimation moyenne du paramètre pour les n_h paramètres estimés,

$$\hat{\theta}_{(h\bullet)} = \sum_{i=1}^{n_h} \hat{\theta}_{(hi)} \,/\, n_h \quad (4)$$

de sorte que chaque suppression $(n–1)$ produise le paramètre $\hat{\theta}_{(hi)}$ de l'équation (4), c'est-à-dire une estimation de l'indice total du prix des produits de consommation courante de l'équation (3), $\hat{\theta} = I^{0,t}$, avec le $i^{\text{ème}}$ article supprimé.

L'estimateur de variance *jackknife* de l'indice finalement calculé pour tous les groupes de produits dans le secteur des produits de consommation courante est :

$$v\left(\hat{\theta}\right) = \sum_{h=1}^{L} \frac{w_h}{n_h} \sum_{i=1}^{n_h}\left(\hat{\theta}_{(hi)} - \hat{\theta}_{(h\bullet)}\right)^2 \quad (5)$$

Notons que w_h dans l'équation (5) est un facteur de correction en fonction des strates, $w_h = \left(n_h - 1\right)\left(1 - \frac{n_h}{N_h}\right)$, sans échantillonnage de remplacement.

Résultats de l'estimation

Sur la base de n = 2 066 cycles des L = 231 strates complètes (n > 1), l'écart-type estimé de la variation de l'indice des prix des produits de consommation courante calculé avec des données de caisse est de 0.168 unités indicielles en moyenne sur les douze mois de l'année 2016, c'est-à-dire la variation mensuelle par rapport à la période de base. Ainsi, pour une valeur de l'indice de 102, l'incertitude d'un intervalle de confiance à 95 % devient [101.67 ; 102.33]. Les estimations de l'écart-type mensuel sont indiquées au tableau 3.

Les résultats du tableau 3 méritent d'être remis dans le contexte de la réalité pratique. Si les échantillons découlaient d'un échantillonnage aléatoire simple et si, au même moment, la consommation de biens était également distribuée entre tous les produits au sein de chaque groupe de produits (c'est-à-dire si les préférences des consommateurs étaient identiquement hétérogènes et la consommation répartie à parts égales sur tous les articles), les résultats obtenus pourraient facilement être étendus à l'ensemble de tous les produits. Dans

8. C'est actuellement le cas pour Statistics Sweden. D'autres options sont possibles : CBS (Pays-Bas) applique le calcul de l'indice, les agrégats élémentaires, aux enseignes individuelles, ce qui représente un niveau un peu plus détaillé qu'ici (van der Grient & de Haan, 2010).

9. En réalité, certains produits ont des poids individuels reflétant de gros volumes de consommation. Cela est ignoré ici afin d'éviter toute volatilité dans les estimations de variance qui ne serait due qu'à la pondération. Tous les produits de l'échantillon sont considérés comme le résultat d'un échantillonnage aléatoire simple.

Tableau 3
Estimations de l'écart-type

Mois en 2016	écart-type
Janvier	0.1725
Février	0.1464
Mars	0.1514
Avril	0.1668
Mai	0.1692
Juin	0.1705
Juillet	0.1825
Août	0.2047
Septembre	0.1651
Octobre	0.1684
Novembre	0.1805
Décembre	0.1426

Note : valeur en unités indicielles. Indice des produits de première nécessité basé sur des données de caisse. 2 066 produits et 231 strates.

un tel cas hypothétique, et sachant qu'un magasin de produits de consommation courante vend en moyenne plus de 10 000 articles, l'échantillon utilisé pour l'IPC suédois (800 articles) permettrait, avec une couverture de 8 % de calculer la variance totale, *via* l'ajustement pour une population finie, $(1 - (n/N))$. Si la taille de l'échantillon est $n = 800$ et que la taille de la population est $N = 10\,000$, la correction en population finie est de $(1 - (800/10\,000))$, appliquée au tableau 3.

Les écarts-type estimés peuvent être évalués relativement à l'écart-type total de l'IPC. La part des produits de consommation courante dans l'IPC est de 13.9 %, comme indiqué au tableau 1, tandis que l'écart-type total de l'IPC, pour le taux d'inflation annuel, est estimé à 0.12 unités d'indice (SCB, 2017). Si l'écart-type estimé pour les produits de consommation courante est relié à cet écart-type total en fonction de la pondération, alors seulement 4 % de la variance de l'IPC résulte des produits de consommation courante (le poids est au carré, de même que les écarts-type, pour obtenir les niveaux appropriés). En raison de cette faible contribution à la variance, une augmentation de la taille de l'échantillon ne permet pas d'obtenir une précision beaucoup plus importante de l'IPC global, même si les éléments inclus dans le panier découlent d'un échantillonnage aléatoire simple.

La pondération des articles, explicite ou implicite par le biais d'un échantillonnage proportionnel à la taille, compense ce calcul linéaire puisqu'il s'agit d'un effet du plan de sondage. Pour cette raison, prendre un échantillon des quelques articles les plus vendus et de quelques articles représentatifs pour le reste engendre en pratique une contribution à la variance inférieure à celle qui découlerait d'une simple estimation de la variance, comme nous le faisons ici. Une approche alternative consisterait à utiliser le panier dynamique avec une limite pour les articles les plus vendus. Certes, l'application d'un tel seuil en termes de part de valeur par groupe de produits améliorerait la précision, mais cela n'est pas nécessairement mieux pour estimer l'inflation – c'est plus simple, mais probablement seulement un peu plus précis puisque la consommation n'est pas répartie à parts égales entre tous les articles.

Interactions et propriétés de population finie

Il peut y avoir des liens, en termes de niveau de prix, entre les articles et les magasins et, en conséquence, entre les marques. Il peut être pertinent de tenir compte de cette interaction dans le cadre de l'estimation de la variance de l'IPC, comme l'explique Norberg (2004). Toutefois, dans la mesure où on considère ici que l'échantillon de magasins est fixe, toute interaction potentielle est ignorée dans ce qui suit, probablement sans affecter les résultats.

Une autre caractéristique de l'échantillon d'articles existant est sa propriété de population finie. Comme nous l'avons indiqué, les échantillons d'articles sont obtenus à partir de bases de sondage complètes présentant une couverture quasi complète pour l'année concernée, y–2. Dans la mesure où la probabilité d'échantillonnage est proportionnelle à la taille, certains

articles/produits de l'échantillon sont les plus vendus et ont donc une probabilité d'inclusion égale à 1. Par conséquent, la variance effective est inférieure à celle estimée ici, car l'estimation jacknife traite tous les items avec la même probabilité d'inclusion alors qu'elle varie dans la réalité. Le compromis proportionnel correspond au scénario le plus défavorable, comme si tous les articles étaient échantillonnés avec une probabilité égale.

Changements de quantités dans les produits de consommation courante

On examine maintenant certains changements de quantités observés et notables sur le marché suédois des produits de consommation courante, et leur impact sur l'IPC si les produits concernés font partie de l'échantillon utilisé pour l'IPC[10]. Jusqu'à présent, à notre connaissance, les cas de lots dont la taille augmente ou de paquets de plus en plus pleins sont assez rares, et les questions soulevées ici concernent des diminutions de la quantité, qui ont été reprises par les médias suédois. En un sens, la quantité peut être considérée comme un aspect de qualité – d'ailleurs les deux termes sont parfois utilisés de façon interchangeable (voir le Manuel de l'IPC §7.77, ILO, 2004). Lorsque c'est nécessaire, des ajustements de quantité sont effectués pour les articles (de remplacement) entrants afin que leur prix soit exprimé pour des unités comparables à celles des produits précédents (ceux de la période de base).

Substitution d'articles et ajustements en unités comparables

Le plan de sondage et l'introduction d'articles de remplacements au sein du panier de l'IPC ont un intérêt particulier pour assurer la comparabilité durant l'année, comme souligné dans le Manuel de l'IPC (ILO, 2004, Ch. 8) qui examine aussi les données de caisse. Ainsi, il est noté que : « *Si peu de choses changent en termes de qualité et de gamme des produits disponibles, la méthode des modèles appariés présente de nombreux avantages. La méthode des modèles appariés permet de faire des comparaisons à périmètre constant dans des magasins semblables.* [...] *En cas de roulement très rapide des articles engendrant un épuisement rapide de l'échantillon, les remplacements ne sont pas fiables pour maintenir l'échantillon. Des mécanismes alternatifs, qui échantillonnent le double univers des articles dans chaque période ou qui utilisent ce double univers, doivent être employés. Citons par exemple les indices chaînés et les indices hédoniques* [...] » (*ibid.,* § 8.62).

Il apparait clairement que, en cas d'attrition au sein du panier, ou plus précisément de diminution de sa représentativité, une procédure de mise à jour plus rapide du chaînage et du rééchantillonnage mensuel sera probablement plus efficace et plus appropriée pour les données de caisse. Mais reste la question des changements de quantités des produits : le début de la citation ci-dessus (« *Si peu de choses changent...* ») suggère que la méthode des modèles appariés ne fonctionne pas. La principale différence entre la formule d'un indice chaîné mensuellement et celle du panier fixe repose sur le fait que les changements de quantité, s'ils ne sont pas traités, affectent le panier fixe en fonction du calendrier – le biais dépend du nombre de mois restants jusqu'à la mise à jour annuelle de l'échantillon. Une procédure de chaînage mensuel ne fait que renvoyer le problème à plus tard.

La question connexe est celle des valeurs unitaires. Elle est examinée par von Auer (2011) qui discute des indices basés sur la valeur unitaire lorsque les produits sont similaires mais pas identiques, et des valeurs unitaires dans la durée. L'un des critères de similitude est la taille du lot – la commensurabilité – pour laquelle une stratégie de « valeur unitaire modifiée » est présentée. La valeur unitaire modifiée consiste à transformer/recalculer, de façon linéaire, la taille des lots en unités communes entre les produits similaires, afin de préserver la comparabilité avec la période de base[11]. Bien que cela ne puisse pas être utilisé directement dans notre analyse, cette approche semblerait très pertinente : les vraies valeurs unitaires sont en quelque sorte reportées à la période de base. Cette approche revient à un panier, et non simplement l'indice, basé sur la valeur unitaire. Ici encore, l'enjeu méthodologique est de pouvoir faire des comparaisons pertinentes et non pas de contourner le problème[12]. En particulier, qu'il s'agisse du concept de changement des niveaux de prix ou de la méthodologie conventionnelle de l'IPC, la linéarité du calcul de la valeur unitaire proportionnelle peut être mise en question. Des études menées en interne par Statistics Sweden

10. Pour des raisons de confidentialité, les produits spécifiques inclus dans le panier de l'IPC ne peuvent pas être communiqués. Toutefois, ces exemples sont de source publique et concernent ici l'impact potentiel de situations hypothétiques sur l'IPC.
11. von Auer (2011) s'intéresse au changement des niveaux de prix, notion qui s'éloigne de la moyenne des changements de prix de l'IPC.
12. Le chaînage et l'orgueil démesuré des statisticiens des prix ont fait l'objet d'une étude approfondie par feu le Professeur Peter von der Lippe. cf. www.von-der-lippe.org (19/07/2017).

montrent que la corrélation taille/prix n'est pas proportionnelle mais plutôt exponentielle, en-dessous du niveau unitaire, c'est-à-dire qu'un doublement de la taille engendre moins d'un doublement du prix.

Changements des quantités : quelques exemples sur le marché suédois

Ces dernières années, plusieurs changements d'emballage de certains produits de consommation courante sont intervenus sur le marché suédois. Certains de ces changements ont directement affecté le calcul de l'IPC en raison des ajustements de quantité correspondants aux prix de la période de base pour le panier fixe. Sans ces ajustements, il pourrait se produire un biais notable dans l'IPC en termes d'inflation cachée. Les exemples suivants en fournissent une illustration.

Le café : de nombreux paquets de café ont diminué en taille, passant d'une contenance auparavant « standard » de 500 grammes à maintenant 450 grammes, soit une baisse de 10 %. D'ailleurs, la plupart des paquets actuellement vendus contiennent moins de 500 grammes. Le prix du café peut être assez fluctuant et les ventes par lot sont courantes, par exemple « trois pour le prix de deux », faisant de cet article un cas intéressant. La variation de 10 % de la taille du paquet a été prise en compte manuellement, suivant la procédure standard de l'IPC. Toutefois, la question du changement réel de prix est discutable. De fait, la hausse de prix implicite suite au changement de la taille des paquets a fait la une des médias lors d'un conflit entre un grand distributeur suédois de produits de consommation courante et un producteur de café détenant une part importante du marché[13]. Avec un chaînage mensuel, ce changement ne serait pas détecté. Le poids du groupe de produit « Café » s'élève à 0.39 % dans le panier, de sorte que, si aucun ajustement n'était fait, une inflation de 0.039 (découlant du changement de 0.1 de la taille unitaire) serait ignorée. À noter toutefois qu'elle serait sans doute amalgamée avec les fluctuations du prix du café.

Le lait caillé : en 2015, au moins un producteur laitier a changé le contenu de la brique de lait caillé (le « *filmjölk* », produit spécifiquement suédois très populaire, proche du yaourt et décliné dans plusieurs variétés et teneurs en matière grasse), en passant des litres aux grammes. Ce changement est passé quasiment inaperçu jusqu'à ce que la presse quotidienne et la radio publique[14] l'annoncent dans un flash info. Dans la mesure où le litre est une mesure de volume et le gramme une mesure de poids, et sachant que la densité d'un produit laitier dépend de sa teneur exacte en matières grasses[15] (FAO, 2012), il était donc difficile d'évaluer la variation de quantité. Des ajustements ont été effectués de façon pragmatique pour toutes les marques et les variétés de l'échantillon utilisé pour l'IPC.

Le groupe de produits correspondant, qui couvre à la fois les yaourts et le lait caillé, représente 0.419 % du panier. Suite à la réduction de quantité de 3 % (une mesure approximative du changement de volume), une brique de 1 000 millilitres ne contient plus que 970 millilitres. Puisque les prix n'ont pas changé au point de vente, cela engendre un biais de 0.03 unité pour plusieurs produits inclus dans l'IPC pour un poids agrégé de 0.419 % du panier. Si au moins un tiers du groupe de produits est constitué par ces produits, le biais se chiffre alors à 0.013 %. Pris de façon isolée, ce chiffre apparait très faible. Mais dans un contexte global, l'addition (ou la multiplication) de biais pour divers les articles peut devenir conséquente au fil du temps, et modifier l'évolution de l'indice.

Le tabac : en raison de la réglementation européenne, la taille des paquets des produits inclus dans le groupe du tabac (qui comprend les cigarettes et le « *snus* », tabac humide spécifique à la Suède) a changé. Le nombre de cigarettes dans un paquet a oscillé entre 19 et 20 cigarettes. Ces changements de quantité doivent être comptabilisés dans le panier fixe lors des remplacements, sans quoi les prix en vigueur sur le marché ne correspondent plus à la même quantité de produit, et ce changement de 0.05 unité engendre un biais sur les articles de tabac dû aux cigarettes. Le poids des produits de tabac est de 1.545 % au sein du panier, et celui des cigarettes de 1.01, soit un biais de 0.05 uniquement dû aux cigarettes.

Au total, si les biais présentés dans ces trois exemples étaient masqués dans le chaînage, un biais total d'environ 0.1 % pourrait survenir ($\approx 0.039 + 0.013 + 0.05$ % des poids). Cela peut être comparé à l'écart-type de 0.168 unité d'indice dans le cas de l'échantillonnage aléatoire simple, soit une surestimation de l'écart-type effectif.

---

13. Cela a suscité une petite controverse, un producteur de café suédois estimant que les prix imposés aux consommateurs reflètent la politique de prix des détaillants et non pas celle des producteurs (Berge, 2016).
14. Voir l'expérience réalisée par la radio publique en Suède (Sveriges Radio) dans Bressler & Näslund (2015).
15. Des données sont disponibles sur Internet pour calculer la densité du lait. Nous n'avons pas de chiffre exact pour ce produit suédois spécifique.

* *
*

Avec l'arrivée de nouvelles sources de données, de nouvelles possibilités voient le jour. La couverture, une caractéristique des données massives telles les données de transactions, est incontestable en termes de contexte et de portée. Ces données sont de l'ordre des recensements, moins d'un siècle après l'introduction de la théorie de l'échantillonnage aléatoire, qui visait à préserver la représentativité par le biais d'échantillons petits et rentables (sur la théorie de l'échantillonnage aléatoire, voir Neyman, 1934 ; plus généralement sur les enquêtes par sondage, voir la passionnante anthologie de Betlehem, 2009).

Les données de caisse ont quelque peu mis en question la méthodologie traditionnelle de production de l'IPC, en particulier avec le développement de nouvelles méthodes empruntées à l'analyse des « big data » (par exemple l'apprentissage automatique). De ce point de vue, Statistics Sweden a avancé avec prudence, dans un premier temps à petite échelle, pour préserver la comparabilité dans le temps et avec d'autres pays aux fins de l'harmonisation des indices des prix à la consommation, et pour assurer la transparence.

Dans cet article, nous nous sommes concentrés sur l'inclusion dans l'IPC des données de caisse pour les produits quotidiens, et particulièrement sur la question de l'arbitrage entre la variance liée à l'item et le biais lié au fait de ne pas tenir compte explicitement des ajustements de quantités. Une hypothèse implicite est l'absence de changement technologique, c'est-à-dire que les développements technologiques n'ont pas d'impact direct sur les prix des aliments et des boissons à court terme, de sorte que l'approche traditionnelle du panier fixe peut être maintenue toute l'année. De plus, la collecte manuelle des prix demeure le moyen le plus courant de produire l'IPC, y compris avec des comparaisons directes et des ajustements quantitatifs en cas de remplacement d'un article. Nous avons vu, pour les produits de consommation courante, que la contribution à la variance ou à l'écart-type d'un élément échantillonné au hasard est plutôt faible et aurait tendance à diminuer avec un échantillonnage approprié. Étant donné que les échantillons sont fondés sur des stratégies d'échantillonnage proportionnelles à la taille, la précision est en fait plus élevée que ne le suggèrent les résultats de cet article – bien qu'inférieure à celle obtenue dans les approches dynamiques couvrant des volumes de ventes plus importants. Cela doit être reconnu comme un avantage des méthodes dynamiques, mais l'ampleur de l'amélioration de la précision n'est pas certaine, notamment en raison de la dépendance entre les produits et les détaillants. Comme le montre l'article, les approches mécaniques sans contrôle peuvent être mises en question, non pas en termes de couverture, mais parce que l'indice qu'elles génèrent peut masquer l'inflation plutôt que la montrer si les changements de quantité sont ignorés.

Bien que l'article ait traité du cas des produits de consommation quotidienne, ce problème concerne l'IPC global, ce qui met en lumière un inconvénient possible de l'utilisation de données de caisse : des détails importants comme les ajustements quantitatifs peuvent maintenant être brouillés dans le déluge de données – comme si la couverture seule était la panacée pour obtenir des mesures précises de l'inflation.

Cependant, cela ne doit pas conduire à ignorer ou à nier les opportunités offertes par les données de caisse. De nombreux développements sont en cours dans d'autres pays, comme en témoignent les meetings du Groupe d'Ottawa, le plus important forum mondial sur les indices des prix. Il convient ici de noter les avancées de Statistics Netherlands (CBS) dans ce domaine, dont témoignent divers rapports de recherche publiés par l'institut néerlandais. Néanmoins, d'un point de vue comparatif, l'utilisation de données de caisse avec des méthodes isolées, qui ne peuvent être comparées mais qui modifient significativement la méthodologie de l'IPC, peut être discutable. L'effort pourrait également être disproportionné pour obtenir une légère augmentation de la précision globale : nous avons vu ici que les variations de prix de produits de nécessité quotidienne (à l'exclusion des fruits et légumes) sont faibles, ce qui peut être comparé à d'autres sources d'erreur qui peuvent affecter l'IPC.

Pour finir, l'arrivée des « big data » devrait nous inviter à garder à l'esprit que la production de statistiques exige une évaluation de la qualité du processus dans son ensemble, et pas seulement des données, comme le soulignent par exemple Biemer *et al*. (2014) et Biemer & Lyberg (2003). Il s'agit de penser en termes d'« erreur totale d'enquête » (Biemer *et al*., 2017). Pour les données de caisse, et en particulier l'échantillonnage dynamique, cela implique un contrôle de qualité au niveau de la codification au sein de la nomenclature COICOP. Sinon, les données pourraient ne pas être cohérentes avec le panier.

S'assurer que les données sont cohérentes avec la méthodologie de l'enquête est une question de précaution, comme le note par exemple Couper (2013), qui souligne la nécessité d'adapter les données au sujet plutôt que de déformer le sujet pour l'adapter aux données.

Alors que d'autres pays ont commencé à mettre en œuvre des approches développées avec les « big data », Statistics Sweden s'est tenu pour l'instant à la méthodologie traditionnelle de l'IPC. Mais d'autres étapes dans l'utilisation des données de caisse sont probables dans un proche avenir. □

BIBLIOGRAPHIE

von Auer, L. (2011). The Generalized Unit Value Index. Universität Trier, *Research Papers in Economics* N° 12/11.

Balk, B. (1989). On calculating the precision of consumer price indices. *Contributed Papers 47th Session of the ISI,* Paris.

Balk, B. (1991). Estimating the precision of a consumer price index: some experiences from the Netherlands. *Contributed Papers 48th Session of the ISI*, Cairo. Also presented at the Joint ECE/ILO Meeting on Consumer Price Indices, 18-21 November, Geneva. Modified version in *Netherlands Official Statistics,* 7(1), 48–49.

Bäckström, U. (1997). What Lessons Can be Learned from Recent Financial Crises? The Swedish Experience. *Speech at the Federal Reserve Symposium* Jackson Hole, Wyoming, USA, August 29, 1997. www.riksbank.se/pagefolders/1722/970829e.pdf

Berge, A. (2016). Viktfiffel. *Råd & Rön*, 19 April 2016. www.radron.se/artiklar/viktfiffel/ (accessed on July 26th 2017)

Betlehem, J. (2009). The rise of survey sampling. Statistics Netherlands, *Discussion paper* N° 09015. https://hdl.handle.net/11245/1.312955

Biemer, P., Trewin, D., Bergdahl, H. & Japec, L. (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics*, 30(3), 381–415. https://doi.org/10.2478/jos-2014-0022

Biemer, P. P. & Lyberg, L. E. (2003). *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, C. N. & West, B. T., Eds (2017). *Total Survey Error in Practice*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Bilius, Å., Bubuioc, R. & Tongur, C. (2017). Bestämning av prisvariabeln vid utökad användning av kassaregisterdata för viktvaror. Paper prepared for the CPI Board at Statistics Sweden. https://www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/3/3-1-bestamning-av-prisvariabeln-for-viktvaror.pdf

Boskin, M. J, Dulberger, E. R, Gordon, R. J, Grilliches, Z. & Jorgenson, D. W. (1997). The CPI Commission: Findings and Recommendations. *The American Economic Review*, 87(2), 78–83. https://www.jstor.org/stable/i352631

Bressler, P. & Näslund, N. (2015). Här mäter P4 Kalmar filjmjölken – bara 9 dl i paketet. Sveriges Radio (Swedish National Radio), 11 May 2015. sverigesradio.se/sida/artikel.aspx?programid=86&artikel=6162534 (accessed on July 26th 2017)

Couper, M. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Notes from talk before ESRA, European Survey Research Association in Ljubljana,* Slovenia, July 15-19, 2013. www.europeansurveyresearch.org/sites/default/files/files/Couper%20keynote.pdf

Dalén, J. & Ohlsson, E. (1995). Variance Estimation in the Swedish Consumer Price Index. *Journal of Business & Economic Statistics*, 13(3), 347–356. https://www.jstor.org/stable/1392194

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

Eurostat (2017a). *Practical Guide for Processing Supermarket Scanner Data*.

Eurostat (2017b). *HICP Methodological Manual*.

Englund, P. (2015). The Swedish 1990s banking crisis. A revisit in the light of recent experience. Paper for the *Riksbank Macroprudential Conference,* Stockholm 23-24 June 2015.
www.riksbank.se/Documents/Avdelningar/AFS/2015/Session%201%20-%20Englund.pdf

Food and Agriculture Organization of the United Nations (2012). FAO/INFOODS Density Database Version 2.0 (2012), prepared by Charrondiere, R. U., Haytowitz, D. & Stadlmayr, B.
http://www.fao.org/docrep/017/ap815e/ap815e.pdf

Fisher, I. (1922). *The Making of Index Numbers*. Boston, MA: Houghton-Mifflin.

van der Grient, H. & de Haan, J. (2010). The use of supermarket scanner data in the Dutch CPI.
www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2010/zip.6.e.pdf

van der Grient, H. & de Haan, J. (2011). Scanner Data Price Indexes: The "Dutch Method" versus Rolling Year GEKS.
http://m.stats.govt.nz/ottawa-group-2011/~/media/Statistics/ottawa-group-2011/Ottawa-2011-Presentations/deHaan-2011-presentation-Dutch-scanner-method.pdf

Horrigan, M. W. (2013). *Big Data: A Perspective from the BLS. Amstat News*.
magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/ (accessed on July 7th 2017)

HUI Research (2017). Dagligvarukartan 2016. Handelns Utredningsinstitut.
www.hui.se/statistik-rapporter/index-och-barometrar/dagligvarukartan (accessed on July 7th 2017)

Hull, I., Löf, M. & Tibblin, M. (2017). Webbinsamlade prisuppgifter och kortsiktiga inflationsprognoser. *Ekonomisk kommentar*, Sveriges Riksbank.
www.riksbank.se/Documents/Rapporter/Ekonomiska_kommentarer/2017/rap_ek_kom_nr2_170609_sve.pdf

ILO, IMF, OECD, UNECE, Eurostat, The World Bank (2004). *Consumer price index manual: Theory and practice*. Geneva: International Labour Office.

Johansen, I. & Nygaard, R. (2011). Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages. Paper written for the 2011 *Ottawa Group Conference,* Wellington, New Zealand, 2011.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/b1ab2e631d34d9bbca2578a7007fa493/$-FILE/2011%2012th%20meeting%20-%20Ingvild%20Johansen,%20Ragnhild%20Nygaard%20(Statistics%20Norway)%20Dealing%20with%20bias%20in%20the%20Norwegian%20superlative%20price%20index%20of%20food%20and%20non-alcoholic%20beverages.pdf

Leaver, S. G. & Larson, W. E. (2001). Estimating Variances for a Scanner-Based Consumer Price Index. Bureau of Labor Statistics.
https://www.bls.gov/osmr/research-papers/2001/st010130.htm

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, 558–625.
https://doi.org//10.2307/2342192

Norberg, A. (2004). Comparison of Variance Estimators for the Consumer Price Index. Paper presented at the *8th Ottawa Group Meeting,* Helsinki 2004.
http://www.stat.fi/og2004/norbergp.pdf

Norberg, A., Sammar, M. & Tongur, C. (2011). A Study on Scanner Data in the Swedish Consumer Price Index. Paper presented at the *Twelfth meeting of the Ottawa Group,* Wellington, New Zeeland, 2011.
www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Papers

Norberg, A., Sammar, M. & Tongur, C. (2012). Scanner data – comparability issues. Paper prepared for the CPI Board at Statistics Sweden.
www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/scanner-data-comparability-issues.pdf

Nygaard, R. (2010). Chain Drift in Monthly Chained Superlative Price Index. UNECE, Geneva 2010.
www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2010/zip.7.e.pdf

Ohlsson, E. (1990). Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer Price Index. Statistics Sweden, *R&D Report* N° 1990/6.
https://www.scb.se/contentassets/7d78fc7dc1e643729f7e8388cd3adf32/rnd-report-1990-06-green.pdf

Rodriguez, J. & Haraldsen, F. (2005). The use of scanner data in constructing elementary aggregates for food and beverages – ideas and experiences from Statistics Norway. Statistics Norway, *Unpublished report*.

Rosén, B. (2000). A User's guide to Pareto PPS Sampling. Statistics Sweden, *R&D Report* N° 2000/6.
https://www.scb.se/contentassets/14f5e346f4814dd0acd52d10b23286c6/rnd-report-2000-06-green.pdf

Sammar, M. & Norberg, A. (2012). Sammanvägningsmetod över tre veckor för kassaregisterdata i KPI. Paper prepared for the CPI Board at Statistics Sweden. www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/sammanvagningsmetod-over-tre-veckor-for-kassaregisterdata-i-kpi.pdf

SCB (2017). Kvalitetsdeklaration för KPI. Quality declaration for the Swedish CPI. www.scb.se/contentassets/a1e257bb3a574420b9d3f2ff59851c0a/pr0101_kd_2017.pdf

SOU (2012). Vad är officiell statistik? En översyn av statistiksystemet och SCB. *SOU* 2012/83. https://www.regeringen.se/contentassets/3521811df5b34bd0bed672bd5c71c7f0/vad-ar-officiell-statistik-en-oversyn-av-statistiksystemet-och-scb-hela-dokumentet-sou-201283

Tongur, C. & Sandén, B. (2016). Viktvaror från kassaregisterdata. Paper prepared for the CPI Board at Statistics Sweden. https://www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/utokning-av-livsmedel-i-kassaregisterdata.pdf

United Nations, web page (2017). Detailed structure and explanatory notes. COICOP.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer Verlag.

Comparaison des indices de prix des vêtements et des chaussures à partir de données de caisse et de données moissonnées sur le Web

Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data

Antonio G. Chessa* et Robert Griffioen**

Résumé – Pour collecter les prix des biens de consommation, les instituts de statistique envisagent le recours à des données moissonnées sur le Web comme une alternative possible aux données de caisse. Les données de transaction étant rares, il est naturel de questionner la pertinence des données du Web pour le calcul d'indices de prix. On propose ici de comparer les indices de prix obtenus à partir de données du Web ou de données de caisse pour des vêtements et chaussures vendus par un même magasin en ligne. Les prix constatés en caisse et moissonnés sur le Web sont souvent égaux, bien que ceux du Web soient légèrement supérieurs en moyenne. Le nombre de produits dont les prix sont moissonnés sur le Web est très corrélé au nombre de produits vendus. Compte tenu du taux de renouvellement élevé des articles dans le secteur de l'habillement, une méthode multilatérale (celle de Geary-Khamis) a été utilisée pour calculer les indices de prix. Pour 16 catégories de produits, les indices montrent de légers écarts globaux entre les deux sources de données : les indices en glissement annuel ne diffèrent que de 0.3 point de pourcentage au niveau de la nomenclature COICOP (vêtements pour hommes et pour femmes). Reste à savoir si ces résultats prometteurs pour les données moissonnées sur le Web se confirmeront pour d'autres points de vente.

Abstract *– Statistical institutes are considering web scraping of online prices of consumer goods as a feasible alternative to scanner data. The lack of transaction data generates the question whether web scraped data are suited for price index calculation. This article investigates this question by comparing price indices based on web scraped and scanner data for clothing and footwear in the same webshop. Scanner data and web scraped prices are often equal, with the latter being slightly higher on average. Numbers of web scraped product prices and products sold show remarkably high correlations. Given the high churn rates of clothing products, a multilateral method (Geary-Khamis) was used to calculate price indices. For 16 product categories, the indices show small overall differences between the two data sources, with year on year indices differing only by 0.3 percentage point at COICOP level (men's and women's clothing). It remains to be investigated whether such promising results for web scraped data will also be found for other retailers.*

Codes JEL / JEL Classification : C43, E31
Mots-clés : IPC, données de caisse, moissonnage du Web, méthodes multilatérales, méthode Geary-Khamis
Keywords: CPI, scanner data, web scraping, multilateral methods, Geary Khamis method

Rappel :

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes, et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

** Centraal Bureau voor de Statistiek (CBS), équipe IPC (auteur correspondant ag.chessa@cbs.nl)*
*** Centraal Bureau voor de Statistiek, équipe IPC au moment où ces recherches ont été menées*

Les auteurs tiennent à remercier Eurostat pour la subvention qui leur a été accordée pour mener cette recherche.

Reçu le 31 juillet 2017, accepté après révisions le 1er avril 2019
Traduit de la version originale en anglais

Pour citer cet article : Chessa, A. G. & Griffioen, R. (2019). Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data. *Economie et Statistique / Economics and Statistics*, 509, 49–68. https://doi.org/10.24187/ecostat.2019.509.1984

Les données de caisse sont de plus en plus utilisées pour mesurer l'indice des prix à la consommation (IPC). En effet, elles sont quasiment idéales pour remplacer les données des enquêtes classiques, car elles contiennent des données de transaction. Les prix et les dépenses sont connus pour chaque article vendu, grâce au code-barres (le *Global Trade Item Number*, ou GTIN, émis et géré par la société internationale GS1). Les dépenses correspondant à chaque article, obtenues grâce aux données de caisse, peuvent servir à établir des indices de prix pondérés, ce qui leur confère un avantage conséquent sur les données tirées des enquêtes.

En Europe, jusqu'en 2014, quatre instituts nationaux de statistique (INS) utilisaient les données de caisse dans leur IPC. En janvier 2018, ils étaient dix à le faire (voir également Leclair *et al.*, ce numéro). Bien que les INS soient habilités à développer leur propre méthode de traitement des données de caisse et de calcul des indices de prix pour les agrégats élémentaires, il est néanmoins souhaitable que les méthodes utilisées dans différents pays soient comparables pour ces agrégats. Dans cette optique, afin d'encourager les INS à commencer à traiter les données de caisse, Eurostat a émis des directives et établi une liste descriptive des pratiques actuelles (Eurostat, 2017).

La collecte des données de caisse est un processus potentiellement long. Différents facteurs entrent en jeu. Par exemple, il faut identifier les personnes à contacter chez le détaillant, chercher à savoir si le détaillant est disposé à coopérer et le temps qu'il peut consacrer à préparer un jeu de données dans un format que l'institut de statistique pourra utiliser. Dans plusieurs pays, comme par exemple aux Pays-Bas, une loi sur la statistique peut être invoquée pour demander des données de caisse. En revanche, dans les pays dépourvus d'une telle loi, il peut être difficile d'obtenir ce type de données et les INS se concentrent alors sur les données collectées en ligne (par exemple, voir Breton *et al.*, 2016). Le moissonnage du Web, dans le but de collecter les prix en ligne et des informations sur les caractéristiques des articles, est de plus en plus utilisé depuis quelques années (Breton *et al.*, 2016 ; Cavallo, 2016 ; Griffioen & ten Bosch, 2016) et offre de nouvelles perspectives pour la statistique publique. Comme avec les données de caisse, la taille des échantillons peut être considérablement augmentée, et la collecte et le traitement des données peuvent être largement automatisés. La collecte automatisée de données en ligne permet également de réduire la charge administrative liée à la collecte de prix, non seulement pour les INS mais aussi pour les détaillants. Pour cette raison, le remplacement des enquêtes par sondage par une collecte automatisée de données de prix en ligne représente une opportunité pour les instituts de statistique – mais aussi un défi de taille.

Compte tenu de la popularité grandissante du moissonnage du Web, il est important d'envisager les fonctionnalités et les limitations de l'utilisation des prix en ligne pour le calcul des indices de prix. Le moissonnage du Web ne permet de collecter que les prix en ligne, car les dépenses relatives aux articles proposés sur un site Web ne sont évidemment pas disponibles en ligne. Certes, c'est également le cas pour la collecte de prix classique. Toutefois, maintenant que les données de caisse sont disponibles, on sait quantifier l'effet d'informations présentes ou manquantes sur un indice de prix. Par exemple, les indices de prix peuvent largement différer selon que l'on utilise des poids basés sur les dépenses ou des poids égaux pour les produits dans une formule d'indice (Chessa *et al*., 2017)[1].

De telles différences débouchent sur une question importante : le nombre de prix de produits moissonnés sur le Web est-il bien corrélé au nombre de ventes dans les données de caisse ? Si la réponse est oui, les indices de prix exclusivement basés sur les quantités et les prix moissonnés sur le Web sont susceptibles de fournir une bonne approximation des indices de prix basés sur les données de caisse. Bien sûr, le résultat dépend de plusieurs facteurs, comme la politique pratiquée par les magasins en ligne, la conception de leur site Web (par exemple, quels produits sont mis en avant et apparaissent le plus souvent sur un site) et la stratégie de moissonnage (le site entier est-il moissonné, et si oui à quelle fréquence et à quel moment ?). Certes, une comparaison entre des indices de prix basés sur des données de caisse ou basés sur des données moissonnées n'est raisonnable que si les mêmes métadonnées relatives aux articles peuvent être utilisées dans le calcul des indices concernés.

1. Nous utilisons le mot « produit » en tant que concept générique et le mot « article » en référence au GTIN. Un produit est équivalent à un article lorsque les GTIN affichent un taux de renouvellement peu élevé, c'est-à-dire lorsque les assortiments restent stables au fil du temps. Si les assortiments ne sont pas stables, par exemple lorsque les GTIN sont de courte durée en raison de relances, les GTIN doivent être reliés et classés en différents groupes. Les GTIN de chaque groupe ont les mêmes caractéristiques d'article. Nous appelons ces groupes des « produits ». La façon dont les caractéristiques sont sélectionnées, ainsi que le fait pour les GTIN d'être considérés comme des produits ou non, sont des questions complexes qui mériteraient une étude séparée.

Le Centraal Bureau voor de Statistiek (CBS) reçoit des données de caisse de la part d'un grand magasin en ligne néerlandais depuis plusieurs années. En octobre 2012, CBS a commencé à moissonner des prix en ligne et des métadonnées du même magasin. Disposer de données de caisse et de données du Web fournit donc une excellente opportunité de comparer les prix des produits, les quantités et les indices de prix entre ces deux sources de données. Les indices de prix calculés avec des données de caisse peuvent servir de référence pour évaluer l'exactitude des indices de prix calculés avec des données moissonnées. Cet article vise à comparer les indices de prix basés sur ces deux sources de données.

La suite s'articule comme suit. La prochaine section décrit brièvement les informations contenues dans les données de caisse et les données moissonnées sur le Web du magasin en ligne néerlandais. Dans la section suivante, nous décrivons la méthode appliquée aux données de caisse et aux données du Web, que nous appelons la « méthode Q-U » de l'anglais *Quality-adjusted Unit value*, c'est-à-dire la valeur unitaire ajustée en fonction de la qualité. Les indices de prix calculés sur la base des deux sources de données sont ensuite comparés au niveau de la catégorie et de la nomenclature COICOP[2]. Nous présentons enfin les principales conclusions de cette étude, ainsi que quelques suggestions de recherches complémentaires.

Données de caisse et données moissonnées pour un magasin en ligne néerlandais

Durant les premières années du programme de développement du moissonnage de données du Web, qui a débuté il y a plus de cinq ans, CBS s'est concentré sur les vêtements et les chaussures, dans le but de moins utiliser les enquêtes classiques pour ces catégories de produits au sein de son IPC. En conséquence, la comparaison des prix, des quantités et des indices de prix entre les données moissonnées et les données de caisse portera principalement sur les vêtements et les chaussures. Les résultats d'une analyse statistique des quantités et des prix des produits sont également présentés.

Données de caisse

CBS reçoit des données de caisse d'un grand magasin en ligne néerlandais depuis janvier 2011. Le détaillant spécifie et envoie les données une fois par semaine, et cet accord a également été conclu avec d'autres détaillants. Les données de caisse couvrent les transactions effectuées sur la totalité de l'assortiment du magasin. L'assortiment est extrêmement varié : hormis des vêtements et des chaussures, le magasin vend des appareils électroniques, des articles de maison et de jardin, des produits destinés aux loisirs, etc.

Pour chaque article (GTIN), les jeux de données de caisse contiennent les informations suivantes, qui sont communiquées en tant que champs distincts :

- année et semaine des ventes (informations groupées en un seul champ) ;
- GTIN ;
- numéro de l'article (code à 6 chiffres spécifique au détaillant pour chaque article);
- chaîne de texte contenant une (courte) description de l'article ;
- groupe dans lequel l'article est classé par le détaillant ;
- numéro du groupe ;
- nombre d'articles vendus ;
- chiffre d'affaires (dépenses) ;
- nombre d'articles retournés ;
- chiffre d'affaires des articles retournés ;
- TVA.

Depuis la fin 2013, le nombre d'articles retournés et le chiffre d'affaires correspondant sont également inclus dans les données par le détaillant, et sont communiqués chaque semaine depuis mars 2014. La valeur des articles retournés est déduite des champs « nombre d'articles vendus » et « chiffre d'affaires », de sorte que cette somme est une valeur nette. Pour cette raison, les champs « nombre d'articles vendus » et « chiffre d'affaires » peuvent présenter des valeurs négatives si le nombre d'articles retournés et le chiffre d'affaires correspondant sont supérieurs au nombre d'articles initialement vendus et au chiffre d'affaires correspondant.

Données moissonnées sur le Web

Certains types de produits, comme l'habillement, peuvent présenter un taux de renouvellement

2. Cela correspond aux catégories « Vêtements pour hommes » et « Vêtements pour femmes » de la nomenclature COICOP.

élevé. Les nouveaux articles doivent être reliés aux articles sortants de qualité identique ou similaire, afin que les variations de prix « cachées » soient prises en compte lors du calcul des indices de prix. Ce remplacement d'article prend également le nom de « relance ». Les articles peuvent être reliés en fonction de caractéristiques communes. Pour cette raison, il est important que les jeux de données de caisse contiennent des informations sur ces caractéristiques.

Toutefois, les instituts de statistique dépendent de ce que les détaillants peuvent leur communiquer, de sorte que les métadonnées incluses dans les données de caisse peuvent ne pas suffire à relier les articles. Malheureusement, c'est le cas pour les données de caisse du magasin en ligne traité dans cet article (voir plus bas dans cette section). Les instituts de statistique peuvent contacter les détaillants pour leur demander de plus amples informations. Le moissonnage du Web est une alternative intéressante pour compléter les informations relatives aux articles présentes dans les données de caisse.

L'outil de moissonnage du Web élaboré pour le magasin en ligne néerlandais collecte des données chaque jour depuis son lancement le 6 octobre 2012. Les informations suivantes sont collectées pour chaque article :

- année, mois et jour durant laquelle/lequel les données ont été moissonnées (un seul champ) ;

- numéro d'article spécifique au détaillant ;

- description de l'article ;

- nom de la marque ;

- trois niveaux de classification de l'article ;

- prix de l'article ;

- prix habituel de l'article.

La description de l'article collectée sur le Web contient plus d'informations que celle fournie dans les données de caisse (dans celles-ci, la description de l'article est souvent, par exemple, « Pantalons pour hommes »). Par ailleurs, les chaînes de texte moissonnées contiennent le nom de la marque et le contenu du lot (par exemple le nombre d'articles uniques dans un lot comprenant plusieurs articles), et la taille, le tissu et le style sont précisés pour certains vêtements. Le nom de la marque est également indiqué dans un champ distinct.

Sur le site Web, on peut aller au niveau de l'article à partir du menu principal, en suivant deux menus secondaires, de sorte que les articles sont classés en fonction de trois niveaux de groupe. Comme mentionné plus haut, l'assortiment du magasin en ligne est assez varié. Le moissonnage vise en priorité à collecter des informations sur les vêtements et les chaussures. Les trois niveaux de classification de l'article s'appliquant aux vêtements et aux chaussures se résument comme suit :

- le niveau supérieur (menu principal) divise les vêtements et les chaussures en cinq groupes, à savoir « Vêtements pour hommes », « Vêtements pour femmes », « Vêtements pour enfants », « Haut de gamme » et « En promotion ». Dans cet article, nous appelons ce niveau supérieur le « groupe principal » ;

- le niveau intermédiaire est appelé « catégorie ». L'outil de moissonnage a collecté des informations relatives à 145 catégories durant la période couverte par cette étude (c'est-à-dire de mars 2014 à décembre 2016) ;

- le niveau le plus détaillé est appelé « type », et contient 1 131 groupes.

Les groupes principaux « Haut de gamme » et « En promotion » peuvent contenir des articles dont le prix est réduit. En conséquence, un article peut être atteint à partir du groupe principal « En promotion » ou à partir de l'un des trois groupes principaux « Vêtements pour hommes », « Vêtements pour femmes » ou « Vêtements pour enfants ». L'outil de moissonnage « navigue » parmi ces cinq groupes principaux, de sorte que chaque article peut être moissonné plus d'une fois par jour. Les articles moissonnés plus d'une fois sont comptabilisés à chaque fois.

Il va de soi que le groupe « En promotion » ne contient pas seulement des vêtements et des chaussures mais aussi d'autres articles en promotion. Pour cette raison, l'outil de moissonnage collecte également les informations susmentionnées pour les appareils électroniques, les articles de maison et de jardin, les produits de beauté et de soin, etc. Les données moissonnées contiennent deux prix pour les articles en promotion : le prix réduit de l'article et son prix habituel. Le prix habituel d'un article en promotion est collecté en même temps que le prix réduit ; il correspond au prix en vigueur juste avant la période de promotion. Pour nos calculs d'indices, nous utilisons bien sûr les prix réduits – et non pas les prix habituels – pour les articles en promotion.

Analyse statistique des données de caisse et du Web

Dans cette sous-section, nous analysons plusieurs aspects des données de caisse et des données moissonnées ayant un impact direct sur le calcul des indices de prix. Notre priorité est de comparer les prix calculés à partir de ces deux sources. Les quantités vendues servent à calculer la valeur unitaire des produits et, associées aux prix, permettent d'établir le poids des produits. Dans ce contexte, il convient donc également de se demander comment les quantités vendues peuvent être comparées aux nombres de prix de produits moissonnés sur le Web.

Propriétés des deux jeux de données

Avant d'utiliser de grands jeux de données numériques dans l'IPC ou à des fins de recherche, une première étape clé consiste à effectuer plusieurs vérifications. Les articles sur la qualité des données publiés par Daas & van Nederpelt (2010) et Daas & Ossen (2010) proposent plusieurs « dimensions de qualité » pouvant servir à vérifier les données. Nous résumons ci-dessous nos conclusions sur certaines des dimensions que nous avons analysées dans le cadre des données de caisse et des données du Web.

- Exhaustivité : les variables (c'est-à-dire les colonnes ou les champs) des deux jeux de données présentent un degré d'exhaustivité élevé. Tous les enregistrements des données de caisse sont consignés, à l'exception du code GTIN qui présente un pourcentage élevé de valeurs manquantes (46.4 %). La raison de ce grand nombre de valeurs manquantes est inconnue, mais pourrait être liée au fait que le détaillant a ses propres codes produit, qui sont disponibles pour chaque enregistrement. Une description du produit est également disponible pour chaque enregistrement. Les données moissonnées sur le Web présentent, elles aussi, un degré d'exhaustivité élevé. Le prix et la description du produit manquent dans 21 enregistrements, ce qui est négligeable sur un total de plusieurs millions d'enregistrements.

- Stabilité : la stabilité est un autre facteur essentiel devant être vérifié avant d'utiliser un jeu de données à des fins de production statistique régulière. La production de l'IPC est entravée si, au cours d'un mois donné, le nombre total d'enregistrements semble largement inférieur à la normale. Ni les données de caisse ni les données moissonnées ne reflètent les augmentations ou diminutions rapides du nombre total d'enregistrements mensuels. Le nombre d'enregistrements augmente au fil du temps, ce qui s'explique par l'accroissement de l'assortiment.

- Niveau de détail : le volume de métadonnées incluses dans les données de caisse du magasin en ligne est limité ; ainsi, à titre indicatif 25 % des descriptions d'articles ne contiennent qu'un seul mot et 62 % ne comprennent pas plus de deux mots.

L'outil de moissonnage a collecté des informations relatives à 385 833 articles entre mars 2014 et décembre 2016. Ce chiffre est proche de celui de 407 253 articles vendus indiqué dans les données de caisse, bien que ces dernières couvrent la totalité de l'assortiment (contrairement aux données moissonnées sur le Web). Le grand nombre d'articles moissonnés découle, d'une part, du fait que l'outil de moissonnage collecte également des informations sur des articles ne faisant pas partie du secteur de l'habillement dans les groupes « Haut de gamme » et « En promotion » et, d'autre part, du fait que le site Web peut également comporter des articles non vendus.

En combinant le nom de la marque avec les trois niveaux de classification de l'article afin de regrouper ou relier les articles, les 385 833 articles moissonnés sur le Web se retrouvent divisés en 59 588 groupes d'articles. Le rapport nombre d'articles/groupes d'articles est donc assez faible. Il est beaucoup plus faible qu'avec les données de caisse (1 635 groupes pour 407 253 articles), ce qui indique le plus grand niveau de détail des métadonnées collectées par l'outil de moissonnage. Cela favorise l'homogénéité des produits lorsque les caractéristiques du produit sont utilisées pour définir le produit.

- Respect des délais : CBS reçoit des données de caisse une fois par semaine, habituellement dans les délais impartis, de tous les détaillants. L'outil de moissonnage du Web collecte des données une fois par jour, au cours de la nuit afin de pas occasionner de gêne durant les heures de pointe sur le site du magasin. Les données sont disponibles dès qu'elles ont été collectées. Toutefois, dans certaines circonstances, les délais peuvent ne pas être respectés, par exemple, si un site Web est indisponible ou a changé. À notre connaissance, les sites Web sont rarement indisponibles. En revanche, les changements de site sont plus fréquents, et c'est pourquoi nous avons créé une équipe « DevOps » dédiée au développement et aux opérations, afin

d'adapter l'outil de moissonnage et d'assurer son fonctionnement continu (pour plus de détails sur sa mise en œuvre à CBS, voir Griffioen *et al.,* 2016).

Comparaisons de prix

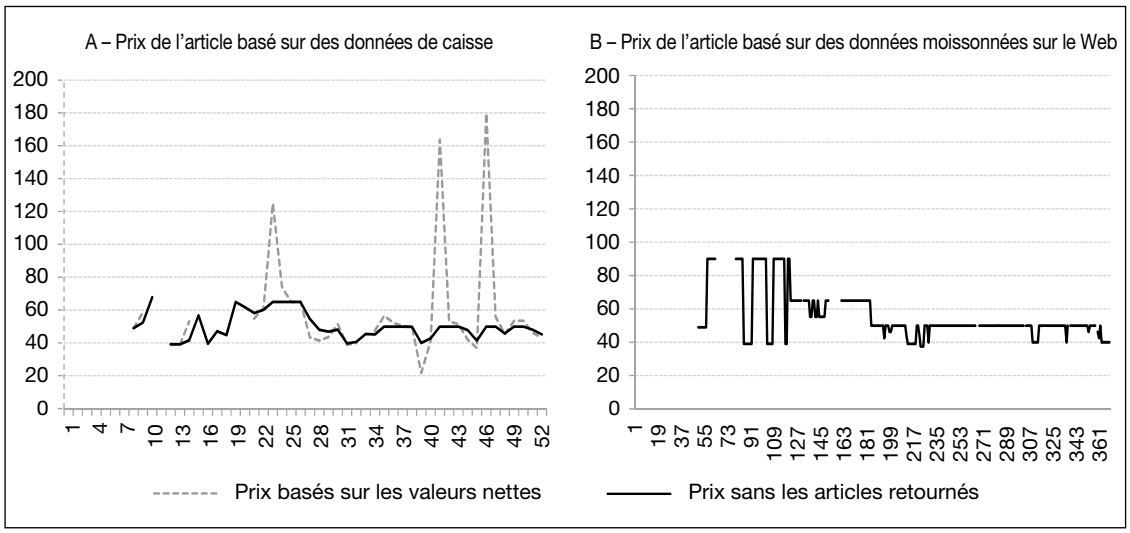
Il convient de noter que les données de caisse, d'une part, permettent de calculer les prix des transactions (c'est-à-dire les prix réellement payés par les consommateurs) et, d'autre part, peuvent avoir des composantes différentes, comme par exemple des réductions spéciales pour les titulaires de cartes ou de bons de réduction. Cela n'est pas le cas pour les prix moissonnés sur le Web, qui ne sont pas les prix des transactions mais les prix offerts par le détaillant sur son site Web.

Le prix d'un ensemble de transactions différentes sur un même article, ou sur des articles de même qualité, peut être calculé en tant que valeur unitaire : il s'agit du ratio des dépenses totales divisées par la somme des quantités vendues (ILO *et al.*, 2004, p. xxii). Ce calcul est habituellement très simple, mais des complications peuvent néanmoins survenir si les consommateurs retournent fréquemment certains articles. Le magasin en ligne propose une politique de retour favorable à ses clients, leur permettant de renvoyer leurs articles gratuitement dans un délai de 14 jours à compter de la livraison.

Les quantités retournées et les dépenses correspondantes sont déduites des quantités vendues et des dépenses durant la semaine pendant laquelle les articles sont retournés et traités par un détaillant. Les quantités vendues et les dépenses représentent donc une valeur nette dans les données de caisse. La semaine durant laquelle le retour est traité peut être différente de celle durant laquelle l'article a été acheté. Cela a deux conséquences importantes : premièrement, les dépenses et les quantités nettes peuvent être négatives ; deuxièmement, la valeur unitaire tirée des deux valeurs nettes diffère du prix initialement payé si le prix d'achat de l'article est différent du prix en vigueur pendant la semaine durant laquelle l'article est retourné. En outre, les consommateurs tendent à acheter plus d'un article lorsqu'il est en promotion. Pour cette raison, les premières semaines suivant une promotion doivent faire l'objet d'une attention particulière lors de la comparaison des prix basés sur les données de caisse avec ceux basés sur les données moissonnées sur le Web.

Lors de toute demande de données de caisse, CBS demande des informations distinctes sur les quantités retournées et sur les dépenses correspondantes. Les données de caisse du grand magasin en ligne néerlandais contiennent ces informations depuis la 12e semaine de l'année 2014. Nous pouvons donc quantifier l'impact des retours d'articles sur les dépenses nettes, les quantités vendues et les valeurs unitaires.

Figure I
Prix hebdomadaires basés sur des données de caisse et prix quotidiens basés sur des données du Web pour un même article (« jeans pour hommes ») en 2015



Note : deux calculs de prix sont indiqués pour les données de caisse : avec les articles retournés (c'est-à-dire sur la base des valeurs nettes) et sans ces articles retournés. Les prix sont exprimés en euros. L'axe horizontal indique le numéro de la semaine (données de caisse) et le jour (données moissonnées sur le Web).
Source : données de caisse pour les prix de vêtements (gauche) et prix moissonnés sur le Web (droite).

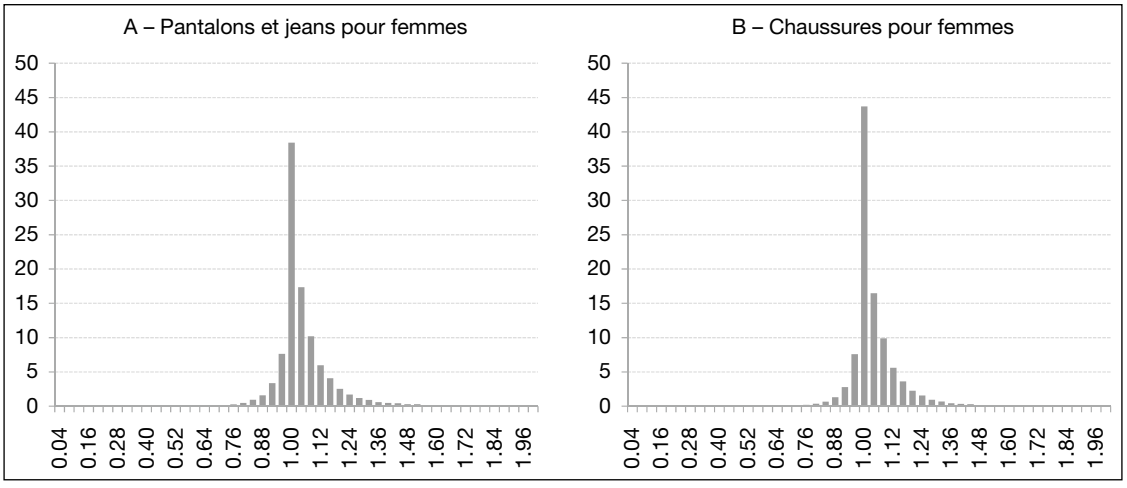
La figure I montre les prix basés sur les données de caisse et ceux basés sur les données du Web pour un même article durant une année entière. Les prix tirés des données de caisse (figure I-A) incluent les retours d'articles, c'est-à-dire que les quantités et les dépenses liées aux articles retournés sont déduites du chiffre d'affaires des semaines durant lesquelles les articles ont été retournés afin de parvenir à la valeur nette. Les prix ont été calculés uniquement si les quantités et les dépenses nettes sont supérieures à zéro. Trois pics importants apparaissent. Chacun de ces pics fait suite à une semaine durant laquelle les prix ont baissé. Les valeurs unitaires calculées à partir des quantités et des dépenses nettes produisent des prix supérieurs aux prix en vigueur pendant la semaine durant laquelle les articles ont été retournés. Les pics de prix surviennent lorsque les quantités d'articles retournés sont proches des quantités vendues pendant la semaine durant laquelle les articles sont retournés.

En soustrayant ces valeurs, on peut calculer le « vrai » prix initial de la transaction (ligne continue sur le graphique de la figure I-A). Cela montre combien il est important de demander des informations séparées sur les dépenses et sur les quantités d'articles retournés. Les prix corrigés sont beaucoup plus proches des prix moissonnés sur le Web indiqués figure I-B. Les prix moissonnés sont plus élevés, en moyenne, durant les premières semaines (c'est-à-dire jusqu'à la 19e semaine ou la 109e journée sur la figure I-B). L'article s'est vendu pour la première fois durant la 8e semaine de 2015. Il semblerait qu'il ait été inclus dans l'assortiment à un prix élevé, mais la ligne continue du graphique I-A suggère que les consommateurs l'ont acheté principalement lorsqu'il était en promotion. Après la période initiale, les écarts entre les prix dans les deux jeux de données se réduisent.

Compte tenu de l'impact que les articles retournés peuvent avoir sur les quantités et les dépenses nettes, nous avons décidé d'exclure les articles retournés des quantités et des dépenses pour la comparaison avec les prix et les quantités moissonnés sur le Web. Nous avons calculé deux statistiques de base : d'une part, le rapport entre les prix moissonnés sur le Web et les prix basés sur les données de caisse et, d'autre part, la corrélation entre le nombre de produits vendus et le nombre de prix de produits moissonnés sur le Web au fil du temps. Nous calculons la corrélation car il est difficile de faire une comparaison bijective entre les nombres de produits vendus et les nombres de prix des produits moissonnés sur le Web.

Les histogrammes représentant ces rapports sont indiqués à la figure II pour les catégories combinées de vêtements « Pantalons et jeans » et « Chaussures » pour femmes. Nous avons combiné dans le même groupe les articles ayant la même marque et le type de classification de l'article le plus détaillé. Nous avons fait le même choix pour le calcul de l'indice de prix (*infra*). Les articles des groupes principaux « Haut de gamme » et « En promotion » ont également été inclus afin de tenir

Figure II
Distribution des fréquences des rapports entre les prix de produits moissonnés sur le Web et les valeurs unitaires pour les données de caisse, pour deux catégories de produits



Note : pour chaque graphique, la somme des fréquences est égale à 100 %. Les rapports de prix des axes horizontaux sont centrés sur les valeurs de classe, avec une amplitude de classe de 0.04.
Source : données de caisse et données moissonnées sur le Web pour des vêtements et des chaussures.

compte des prix réduits. Un groupe [marque×type] peut être, par exemple « Shorts en jean » de la marque X. Toute combinaison [marque×type] est appelée « produit » dans cet article.

Les graphiques de la figure II montrent les rapports de prix combinés de tous les produits au cours de chaque mois. Ils présentent des pics élevés aux alentours de 1 (prix égaux) et sont tous les deux orientés vers des rapports supérieurs à 1. Les prix moissonnés sur le Web tendent à être plus élevés, en moyenne, que les prix des transactions. Nous avons déjà constaté la même chose pour les prix d'un même article (cf. figure I). Le plus faible niveau des prix tirés des données de caisse peut s'expliquer par la réorientation des ventes vers des articles moins chers, par exemple lorsque ces articles sont en promotion (« effet quantité »).

Comparaisons de quantités

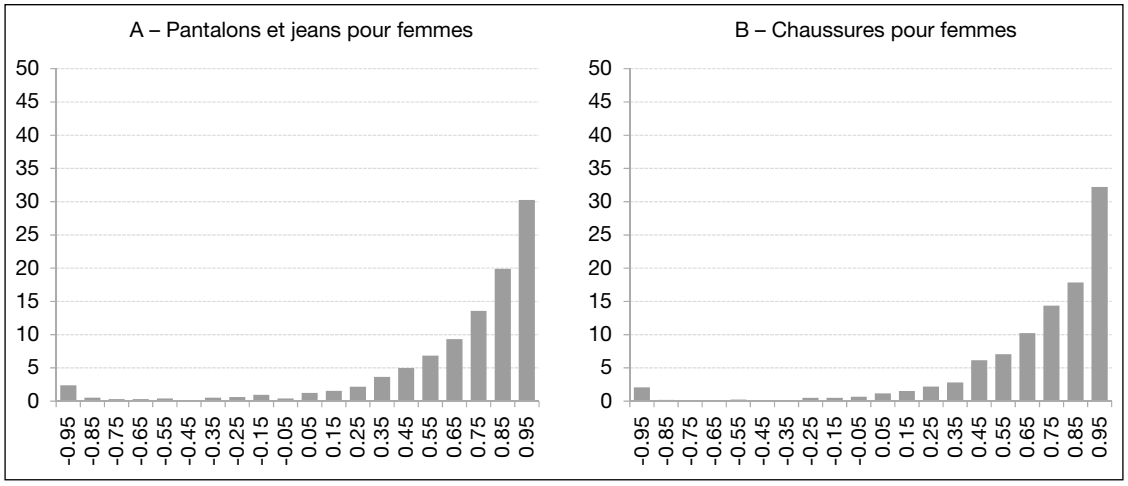
Nous avons calculé la corrélation entre le nombre de produits vendus et le nombre de prix de produits moissonnés sur le Web. Pour chaque produit, nous avons calculé la corrélation entre les couples de nombres vendus et de nombres de prix moissonnés pour tous les mois de la série. Les deux graphiques montrent des corrélations très élevées, et les fréquences les plus importantes s'observent pour les classes où la corrélation est la plus élevée (figure III). Ces tendances n'auraient pas été observées si les nombres moissonnés sur le Web étaient indépendants des nombres de produits vendus, ce qui engendrerait des distributions centrées sur une corrélation nulle. Les légères hausses constatées dans la classe où la corrélation est la plus faible s'expliquent dans une large mesure par des produits dont les prix ne sont observés que pendant deux mois. Si l'on exclut ces produits des calculs, ces hausses disparaissent.

Les fréquences auxquelles les articles peuvent être trouvés dans différents menus d'un site Web au fil du temps semblent correspondre assez bien aux quantités vendues. Cela découle de la politique du détaillant, qui consiste à promouvoir les articles les plus vendus du site. Les autres catégories de produits montrent des résultats semblables, tant en termes de prix que de quantités, ce qui crée des conditions favorables pour comparer les indices de prix basés sur les deux jeux de données. Il est donc important d'être en contact avec le détaillant afin d'obtenir des informations sur sa stratégie d'organisation de son site Web.

Dynamique de l'assortiment

Les vêtements et les chaussures se caractérisent habituellement par un taux de renouvellement élevé. Nous avons analysé la dynamique des assortiments de différentes catégories de produits pour les données de caisse et les données moissonnées sur le Web. Cette dynamique a été quantifiée en

Figure III
Distribution des fréquences des corrélations entre le nombre de prix de produits moissonnés sur le Web et le nombre de produits vendus



Note : pour chaque graphique, la somme des fréquences est égale à 100 %. Les corrélations des axes horizontaux sont centrées sur les valeurs de classe, avec une amplitude de classe de 0.1.
Source : données de caisse et données moissonnées sur le Web pour des vêtements et des chaussures.

introduisant trois mesures : (i) la part des produits vendus ou disponibles durant de longues périodes, c'est-à-dire « le flux », (ii) la part des produits ajoutés à un assortiment durant une année donnée, c'est-à-dire « les flux entrants », et (iii) la part des produits retirés d'un assortiment, c'est-à-dire « les flux sortants ». Nous avons calculé ces trois statistiques de flux en bilatéral, c'est-à-dire pour des groupes de deux mois. Le premier mois, choisi comme mois de référence, est resté fixe. Les produits vendus ou disponibles durant le mois de référence et durant le deuxième mois (mois en cours) sont comptabilisés dans le flux, les produits non vendus/disponibles le mois de base mais vendus/disponibles seulement durant le mois en cours sont comptabilisés dans les flux entrants et les produits disponibles le mois de référence mais non disponibles durant le mois en cours sont comptabilisés dans les flux sortants[3].

Les trois statistiques sont calculées pour chaque mois de la période allant de mars 2014 à décembre 2016, en utilisant respectivement les mois de mars 2014, décembre 2014 et décembre 2015 comme mois de référence. Les statistiques résultent de comptages au niveau du produit, c'est-à-dire pour les groupes [marque×type]. La figure IV montre les trois statistiques de flux pour les « pantalons et jeans » pour hommes.

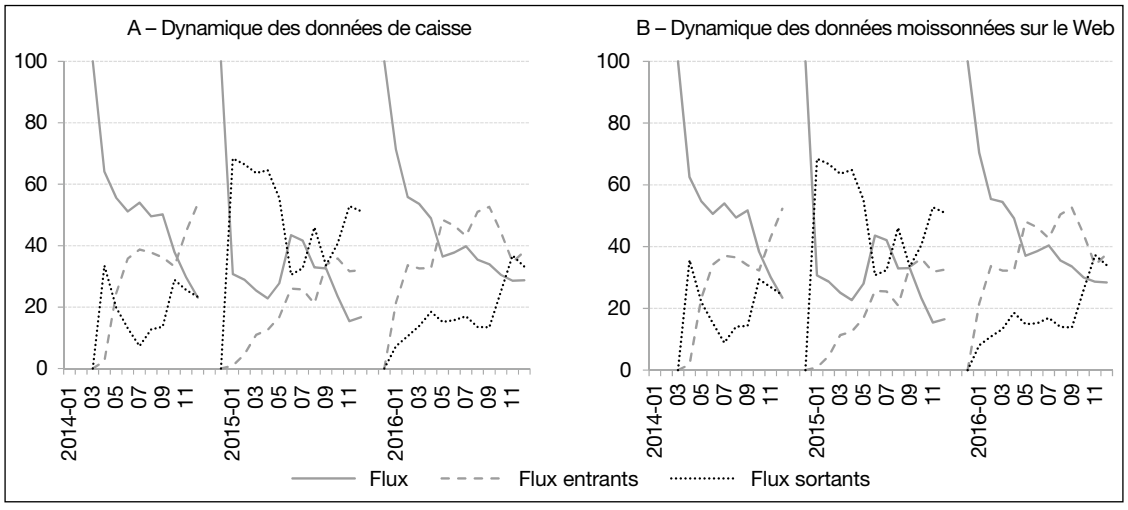
Le taux du flux est, par définition, égal à 100 % pour les mois de référence. La baisse rapide du taux du flux et la hausse, ainsi que le niveau élevé, des flux entrants indiquent un assortiment très dynamique. Les deux graphiques montrent clairement qu'il y a très peu de différence entre les statistiques de flux basées sur les données de caisse et sur les données du Web. Cela signifie que les articles qui ne se vendent plus sont rapidement retirés du site Web. On note également que la forte dynamique s'observe au niveau du produit, c'est-à-dire à un niveau moins détaillé que celui de l'article/du GTIN, ce qui joue un rôle important dans le choix de la méthode d'indice.

La méthode Q-U

Le secteur de l'habillement est notoirement complexe en termes de calcul des indices de prix, car les catégories de produits peuvent présenter des taux de renouvellement élevés. Les méthodes d'indice bilatérales peuvent être problématiques : les méthodes bilatérales directes n'incluent pas les nouveaux produits dans le calcul de l'indice au cours d'une année donnée, mais seulement lors du mois de base suivant, tandis que les méthodes des indices chaînés mensuellement peuvent souffrir du biais de dérive en chaîne. L'étude comparative de Chessa *et al.* (2017) montre que les indices bilatéraux pondérés peuvent varier de manière significative par rapport aux indices transitifs, ce

3. Des mesures bilatérales ont été choisies afin de faciliter les calculs. Il est bien sûr possible de rallonger la période à un plus grand nombre de mois, mais dans ce cas il est plus difficile de caractériser la dynamique. Pour des informations plus détaillées, voir Willenborg (2017).

Figure IV
Dynamique du flux avec les données de caisse et les données moissonnées sur le Web pour les pantalons et jeans pour hommes, par année



Note : les trois mesures du flux sont exprimées en pourcentage et atteignent 100 % chaque mois.
Source : données de caisse et données moissonnées sur le Web pour des vêtements.

qui est contraire à la condition que les méthodes des indices de prix doivent respecter pour éviter la dérive en chaîne.

Contrairement aux méthodes bilatérales, qui utilisent des informations de deux périodes afin de calculer l'indice, les méthodes multilatérales utilisent des informations de plusieurs périodes. L'un des avantages des méthodes multilatérales sur les méthodes bilatérales est que des indices transitifs ne contenant aucun biais de chaînage peuvent être calculés selon différents poids pour différents produits, et peuvent même varier d'un mois à l'autre. Toutefois, certaines méthodes (dont celle dite GEKS, pour Gini-Eltetö-Köves-Szulc) sont sensibles aux biais baissiers pour des assortiments dynamiques dont les produits sortent des prix de liquidation (Chessa *et al.,* 2017). Ces situations ne sont pas rares dans le secteur de l'habillement (Chessa, 2016a). Pour cette raison, nous avons choisi une méthode qui ne souffre pas des problèmes susmentionnés, à savoir la « méthode Q-U » (de l'anglais *Quality-adjusted Unit value*, c'est-à-dire la valeur unitaire ajustée en fonction de la qualité), pour les données de caisse et les données moissonnées du magasin en ligne. Cette méthode a été introduite dans l'IPC néerlandais en janvier 2016 (Chessa, 2016a). Quand on l'utilise pour des comparaisons de prix entre différents pays, elle est appelée « méthode Geary-Khamis » (GK) et représente alors un cas particulier au sein de la classe globale des méthodes Q-U. Pour cette raison, nous préférons le second terme, ou aussi « Q-U–GK ».

Formule d'indice

Chessa *et al.* (2017) comparent les méthodes d'indice bilatérales et multilatérales pondérées et non pondérées sur les jeux de données de caisse de quatre catégories de produits d'un grand magasin néerlandais autre que celui traité dans cet article. Les poids utilisés dans les formules d'indice peuvent engendrer des résultats largement différents de ceux des méthodes basées sur l'équipondération. Cela étant, les poids utilisés dans les méthodes bilatérales peuvent eux aussi être problématiques, notamment lorsqu'ils sont utilisés pour calculer les indices chaînés mensuellement. Ces indices peuvent engendrer un biais de chaînage important, qui découle directement du caractère intransitif des indices bilatéraux chaînés mensuellement.

Les indices bilatéraux directs ne tiennent pas compte des nouveaux produits en temps voulu, ces derniers n'étant inclus que lors du mois de base suivant, sauf si les prix sont imputés pour les mois précédant le mois durant lequel le produit est ajouté à l'assortiment. Une comparaison faite dans le secteur de l'habillement montre que la contribution des nouveaux produits à un indice peut être considérable (Chessa *et al.*, 2017). Les méthodes multilatérales ne souffrent d'aucun biais de chaînage, ce qui permet d'inclure les nouveaux produits en temps voulu et évite d'avoir à imputer les prix.

La dynamique de l'assortiment justifie également de choisir une méthode multilatérale pour les données de caisse et les données moissonnées du magasin en ligne néerlandais. Les écarts constatés dans Chessa *et al.* (2017) entre les indices de prix basés sur différentes méthodes multilatérales ne sont pas grands, mais peuvent avoir une importance significative. La méthode GEKS, ainsi que la méthode CCDI récemment proposée par Diewert & Fox (2017), sont sensibles aux prix de liquidation des articles sortants, ce qui engendre des biais baissiers (Chessa *et al.*, 2017). D'autres méthodes, comme la méthode Q-U et l'indicatrice temps/produit, n'ont pas ce désavantage.

La méthode Q-U a été introduite au sein de l'IPC néerlandais en janvier 2016, et sa première application à l'indice concernait alors les téléphones portables. Depuis janvier 2017, elle est également appliquée aux données de caisse du grand magasin néerlandais susmentionné. La méthode Q-U peut être considérée comme une famille de méthodes, incluant également certaines méthodes bilatérales bien connues comme les indices de Laspeyres, Paasche et Fisher (voir également Auer, 2014). Mais son but premier consiste à établir des indices transitifs multilatéraux. De fait, la méthode élargit le concept de valeur unitaire à des ensembles de biens hétérogènes. Il faut ainsi tenir compte des différences de qualité qui existent entre les produits, raison pour laquelle nous parlons de « valeur unitaire ajustée en fonction de la qualité ». D'autres auteurs, par exemple Auer (2014), parlent de « valeur unitaire généralisée ».

Pour bien expliquer le concept qui étaye la méthode Q-U, introduisons quelques notations. Soit G_0 et G_t des ensembles de produits appartenant à une catégorie de produit G, pour un mois de base 0 et un mois en cours t. Les ensembles de produits de 0 et de t peuvent être différents. Soit $p_{i,t}$ et $q_{i,t}$ les prix et les quantités vendues du produit $i \in G_t$, respectivement, au cours du mois t. Nous voulons identifier des facteurs d'échelle, disons v_i, transformant les prix de différents produits au cours du mois t en « prix ajustés en fonction de la qualité » $p_{i,t} / v_i$. Cette transformation implique

de convertir les quantités vendues $q_{i,t}$ pour chaque produit en quantités $v_i\, q_{i,t}$. À l'expression (3) ci-dessous, les v_i des produits correspondent aux prix déflatés moyens sur un intervalle de temps donné. Les v_i pourraient être interprétés comme des « prix de référence » et les $v_i\, q_{i,t}$ comme des quantités évaluées aux prix de référence des produits.

La transformation des prix et des quantités nous permet de définir et de calculer une « valeur unitaire ajustée en fonction de la qualité » $\tilde{p}_t$ pour un ensemble de produits G_t au cours du mois t :

$$\tilde{p}_t = \frac{\sum_{i\in G_t}\left(p_{i,t} / v_i\right)\left(v_i q_{i,t}\right)}{\sum_{i\in G_t} v_i q_{i,t}} = \frac{\sum_{i\in G_t} p_{i,t} q_{i,t}}{\sum_{i\in G_t} v_i q_{i,t}} \qquad (1)$$

À noter que $\sum_{i\in G_t} p_{i,t} q_{i,t}$, la dépense totale, n'est pas affectée par la transformation.

L'expression (1) peut servir à définir un indice de prix en divisant les valeurs unitaires ajustées en fonction de la qualité en deux mois :

$$P_t = \frac{\tilde{p}_t}{\tilde{p}_0} = \frac{\sum_{i\in G_t} p_{i,t} q_{i,t} \,/\, \sum_{i\in G_0} p_{i,0} q_{i,0}}{\sum_{i\in G_t} v_i q_{i,t} \,/\, \sum_{i\in G_0} v_i q_{i,0}} \qquad (2)$$

Le numérateur à droite de l'expression (2) est un indice mesurant l'évolution du chiffre d'affaires ou des dépenses entre deux mois. Le dénominateur est un indice de quantité pondéré. L'expression (2) montre clairement pourquoi l'indice de prix est transitif : l'indice de chiffre d'affaires comme l'indice de quantité pondéré sont transitifs.

Les poids v_i sont définis comme suit sur un intervalle de temps donné $[0,T]$:

$$v_i = \sum_{z=0}^{T} \frac{q_{i,z}}{\sum_{s=0}^{T} q_{i,s}} \frac{p_{i,z}}{P_z} \qquad (3)$$

L'expression (3) revient à dire que les v_i sont également des valeurs unitaires. Pour chaque produit, les dépenses sont additionnées sur la période concernée $[0,T]$ puis divisées par les quantités vendues pour le produit durant ce même intervalle de temps. Afin d'exclure les variations de prix des v_i et de l'indice de quantité pondéré, les prix en vigueur pour les produits durant des mois différents sont déflatés par l'indice de prix de la catégorie de produit concernée. Les v_i sont également appelés « prix de référence » (ou « prix internationaux » dans un contexte géographique). L'expression (3) représente le choix effectué sur ces prix dans la méthode Geary-Khamis (GK).

Les prix déflatés moyens d'une période donnée sont donc utilisés pour calculer les quantités transformées $v_i\, q_{i,t}$. Les prix des produits en vigueur durant tous les mois d'un intervalle de temps donné $[0,T]$ sont utilisés, comme cela est d'usage, ainsi que dans d'autres méthodes multilatérales. Toutefois, il pourrait être utile d'envisager certaines améliorations pour l'expression (3) : par exemple, les prix réduits pourraient être exclus des v_i afin que les valeurs obtenues représentent mieux la qualité. Cela pourrait faire l'objet de recherches ultérieures.

Le choix des prix pour définir les v_i est courant dans la théorie de l'indice. La méthode Q-U peut être considérée comme une famille de méthodes d'indice, dans la mesure où différents choix effectués pour les v_i donnent lieu à des formules d'indice différentes. Afin de produire plusieurs exemples à titre d'illustration, nous examinons tout simplement l'ensemble de produits vendus durant les deux mois, à savoir $G_0 \cap G_t$. Si nous définissons $v_i = p_{i,0}$ pour chaque produit $i \in G_0 \cap G_t$, alors l'expression (2) devient un indice de prix de Paasche. Si nous définissons $v_i = p_{i,t}$ pour chaque produit i, alors la formule (2) devient un indice de prix de Laspeyres. Si les v_i sont égaux pour tous les produits, alors (2) se simplifie et devient un indice de valeur unitaire. C'est précisément ce à quoi nous pourrions nous attendre pour des produits de même qualité, puisque leurs quantités vendues peuvent être additionnées sans être transformées.

Dans la mesure où l'indice de prix sert de déflateur dans (3), les équations (2) et (3) doivent être résolues simultanément. Chessa (2016a) décrit un algorithme itératif commençant par des valeurs initiales arbitraires pour les indices de prix $P_1,\ldots,P_T$, avec $P_0 = 1$ (voir également Maddison & Rao, 1996). Ces indices de prix sont substitués dans l'expression (3), de sorte que les valeurs initiales puissent être calculées pour chaque v_i. Ces valeurs sont ajoutées à l'expression (2) afin d'actualiser les indices de prix initiaux. Ces deux étapes sont répétées jusqu'à ce que les différences entre les indices de prix des deux dernières étapes d'itération respectent un critère d'arrêt défini par l'utilisateur. Geary (1958), Khamis (1972), Auer (2014) et Chessa (2016a) fournissent des informations plus détaillées sur les méthodes Q-U et GK.

Avant d'appliquer la méthode, plusieurs questions doivent être traitées, et en premier lieu la longueur de la fenêtre temporelle, l'intervalle de temps $[0,T]$, ainsi que la façon dont des données

supplémentaires peuvent être ajoutées à mesure que de nouvelles données deviennent disponibles chaque mois. Nous traitons ensuite la question de la définition des produits inclus dans l'ensemble de biens G_t.

Longueur de la fenêtre temporelle

Pour choisir l'intervalle de temps, nous avons utilisé un mois de base fixe (le mois de décembre de l'année précédente), conformément à la réglementation sur les indices des prix à la consommation harmonisés. L'IPC néerlandais utilise un intervalle de 13 mois, que nous avons aussi retenu ici.

L'impact de tout changement de l'intervalle temporel sur les indices de prix est étudié dans Chessa *et al.* (2017) et, de façon plus poussée, dans Chessa (2017a). La première étude compare les intervalles de 13 mois et la période entière de 50 mois pour quatre catégories de produits. Des différences significatives ont été identifiées pour l'une des catégories. Dans Chessa (2017a), les différences ont également été quantifiées au niveau de la nomenclature COICOP. Les différences entre les intervalles de 13 mois et de 4 ans sont de l'ordre de dixièmes de point de pourcentage pour des indices en glissement annuel, et négligeables dans un nombre assez grand de catégories COICOP. Pour le détaillant de la grande chaîne de supermarchés néerlandaise, il n'apparaît aucune différence entre les deux intervalles.

Actualisation des poids et calcul de l'indice

À mesure que de nouvelles données deviennent disponibles chaque mois, l'inclusion de données supplémentaires peut engendrer des valeurs différentes pour les v_i, et les indices de prix calculés jusqu'au mois précédent peuvent eux aussi changer. Toutefois, sauf circonstances exceptionnelles, il est impossible de réviser les indices de prix dans l'IPC. Compte tenu de ce « problème de révision », comment pouvons-nous calculer un indice de prix pour le mois suivant ?

En théorie, la solution des équations (2) et (3) nous donne un ensemble de 13 indices transitifs pour toute année $[0,T]$, sachant que le mois de base 0 est le mois de décembre de l'année précédente et que $T = 12$ représente le mois de décembre de l'année en cours. Les indices de prix et les poids des produits ou les prix de référence v_i sont calculés pour les 13 mois de l'année de façon simultanée, de sorte que les v_i ont la même valeur chaque mois. Il serait possible de publier les indices qui en découlent s'il était possible de réviser les indices de prix des mois précédents à chaque fois que de nouvelles données sont incluses dans le calcul de l'indice pour un mois ultérieur. Les v_i calculés pour le mois de décembre de l'année en cours finissent par donner les valeurs souhaitées pour les poids des produits, ce qui pourrait servir à produire les indices transitifs chaque mois.

Dans la pratique, comme on ne peut pas prévoir les prix des mois à venir, la construction d'indices transitifs restera, au mieux, une référence théorique idéale. L'inclusion de données d'un mois ultérieur modifie la valeur des v_i et, en conséquence, les indices de prix des mois précédents. Comme en règle générale, il est impossible de réviser les indices de prix de mois précédents dans l'IPC, cela soulève la question du mode de calcul d'un indice de prix pour un mois ultérieur.

Différentes méthodes ont été proposées pour mettre les v_i à jour et calculer les indices de prix d'un mois ultérieur. Ces méthodes sont basées sur des choix relatifs à trois aspects[4] :

- l'utilisation d'un mois de base fixe ou d'un mois de référence mobile ;

- l'adoption d'une fenêtre glissante ou d'un intervalle à expansion mensuelle, sachant que ce dernier ne peut s'utiliser qu'avec un mois de base fixe ;

- l'utilisation d'une méthode d'indice directe, de chaînage mensuel ou de raccordement.

Chessa (2016a) propose une méthode de mois de base fixe, un intervalle à expansion mensuelle et une méthode directe pour calculer l'indice de prix d'un mois ultérieur. Cette méthode utilise des données tirées de nombres de mois différents tout au long d'une année (deux mois en janvier, trois en février et ainsi de suite jusqu'à un nombre maximal de 13 mois en décembre) et ne requiert pas de données historiques. La méthode directe calcule les indices de prix du mois en cours par rapport au mois de base, en utilisant l'ensemble de valeurs le plus récent pour les v_i.

Elle fait en sorte que les indices de prix de décembre correspondent aux indices de prix transitifs qui auraient été obtenus en utilisant les données complètes des 13 mois pour chaque mois de l'année. Ainsi, la méthode de « fenêtre d'expansion mensuelle à base fixe » (ou FBEW de l'anglais *fixed*

4. À noter que ces choix, et en conséquence le type de méthode d'actualisation, peuvent être appliqués en combinaison avec toute méthode multilatérale. Un exemple en est fourni dans Chessa et al. *(2017).*

base monthly expanding window) permet d'éviter le biais de chaînage. Les séries d'indices de plus d'un an sont construites en chaînant la série de l'année en cours à l'indice du mois de décembre de l'année précédente, de sorte qu'une forme de chaînage est finalement réalisée. Mais c'est une forme de chaînage moins fréquente et, par ailleurs, compte tenu de l'utilisation d'un intervalle de 13 mois, la valeur théorique des v_i peut varier d'une année à l'autre pour chaque produit. Il s'agit d'un choix explicite, qui pourrait être fait pour refléter la variation progressive de la qualité au fil du temps.

La fenêtre à expansion mensuelle pourrait également être remplacée par un intervalle glissant de 13 mois, tout en maintenant le calcul des indices de prix à l'aide d'une méthode directe par rapport à un mois de base fixe. Cette méthode alternative est comparée à la méthode FBEW dans Chessa (2017a) et dans Lamboray (2017). Les différences entre les deux méthodes se sont avérées peu importantes ou négligeables. Les indices calculés à l'aide de méthodes d'actualisation et les « indices de référence » transitifs se sont avérés quasiment ou complètement égaux dans chacun des cas étudiés (Chessa, 2016a ; 2017a ; 2017b). Des différences importantes se sont occasionnellement manifestées, pour la plupart sur de courtes durées.

Une autre catégorie de méthodes consiste à utiliser un mois de référence mobile au lieu d'un mois de base fixe. L'un des choix les plus naturels est de combiner un mois de référence mobile avec une fenêtre glissante d'une durée fixe, ce qui permet d'inclure les données d'un mois ultérieur de façon élégante. Différentes méthodes peuvent être envisagées pour calculer un indice de prix pour le mois en cours, que l'on appelle « méthodes de raccordement ». Voir de Haan *et al.* (2016) pour une vue d'ensemble et Chessa *et al.* (2017) et Krsinich (2014) pour des applications précises.

La méthode du « raccordement des variations » chaîne l'indice en glissement mensuel de l'intervalle glissant le plus récent à l'indice du mois précédent, tandis que la méthode du « raccordement des intervalles » de Krsinich (2014) chaîne l'indice en glissement annuel de l'intervalle complet le plus récent à l'indice d'il y a douze mois. La méthode du raccordement des variations est une méthode de chaînage mensuel qui, par nature, subit le biais de chaînage. Bien que la méthode du raccordement relève d'une sorte de méthode directe, c'est aussi une méthode de chaînage à haute fréquence. Les résultats empiriques indiquent un biais de chaînage potentiel, qui pourrait être important (Chessa, 2016b).

Indices de prix pour les données moissonnées sur le Web et les données de caisse

Préparation des données et choix méthodologiques

Nous avons calculé les indices de prix à l'aide de la méthode Q-U pour les vêtements pour hommes et pour femmes du magasin en ligne néerlandais, sur la base des données de caisse et exclusivement à l'aide des données moissonnées sur le Web. Pour établir des comparaisons pertinentes, nous avons complété les données de caisse par les métadonnées tirées des données moissonnées. Pour cela, nous avons relié les deux tableaux de données en utilisant les codes d'article spécifiques au détaillant en guise de clé de couplage. Nous avons calculé les indices de prix de huit catégories de produits dans le segment des vêtements pour hommes (pantalons et jeans, manteaux et vestes, sous-vêtements et pyjamas, chemises, chaussures, vêtements de sport, pulls et gilets, tee-shirts et polos) et dans celui des vêtements pour femmes (pantalons et jeans, manteaux et vestes, robes et jupes, lingerie, chaussures, vêtements de sport, pulls et gilets, tee-shirts et hauts).

Les huit catégories couvrent environ 85 % des dépenses totales en vêtements pour hommes pour la période allant de mars 2014 à décembre 2016, et environ 80 % des dépenses totales en vêtements pour femmes. Les articles en promotion et haut de gamme ont également été inclus[5].

La première étape préalable au calcul de l'indice de prix consiste à définir le produit. Bien que cela ne soit pas la priorité de cette étude (qui vise à comparer les données de caisse et les données moissonnées sur le Web), il est évident que cette démarche doit être effectuée avec soin, puisque les indices peuvent être très sensibles à toute variation du degré de différenciation entre les produits (Chessa, 2016a; 2017b).

Le secteur de l'habillement affiche habituellement un taux de renouvellement élevé, qui se remarque également à un niveau moins détaillé que le niveau de l'article ou du GTIN (voir la figure IV). Les articles sortants et les nouveaux articles de qualité identique ou similaire doivent être reliés afin d'éviter que les indices ne subissent des biais à la baisse, dont l'ampleur peut être extrêmement importante si les articles sortent de l'assortiment

5. Les articles hors secteur de l'habillement inclus dans ces deux groupes ont été exclus durant l'extraction des données pour chacune des catégories susmentionnées.

à des prix de liquidation (Chessa, 2016a). Les articles sortants et les nouveaux articles peuvent être reliés par des caractéristiques communes, ici le nom de la marque et le « type », c'est-à-dire le niveau le plus détaillé de classification d'article.

Les articles sont donc rassemblés dans le même groupe lorsqu'ils appartiennent aux mêmes groupes [marque×type], que nous appelons « produits ». Les produits doivent être homogènes, c'est-à-dire que les articles d'un groupe donné doivent être de qualité identique ou comparable. Cette question devrait être étudiée de façon plus détaillée dans des travaux ultérieurs, notamment lorsque l'on envisage d'inclure les données d'un magasin en ligne dans l'IPC. La taille moyenne des « produits » va de 7 à 16 articles ; sachant que les codes d'article et les GTIN sont habituellement différents pour les différentes tailles de vêtements, et que ces vêtements sont considérés comme étant de même qualité, cette fourchette suggère que les définitions des produits sont assez étroites.

Les choix suivants ont été faits pour appliquer la méthode Q-U aux données de caisse et aux données du Web.

- Pour les données de caisse, les valeurs unitaires ont été calculées pour chaque produit lors de chaque mois durant lequel il a été vendu. Les dépenses et les quantités d'articles vendus pour un produit donné ont été additionnées, et le chiffre d'affaires correspondant aux articles retournés a été exclu.

- Pour les données moissonnées, les prix mensuels moyens ont été calculés pour chaque produit. Les quantités vendues ont été remplacées par le nombre total de prix moissonnés pour un produit donné au cours d'un mois, puis additionnées pour tous les articles. Les articles peuvent être moissonnés plus d'une fois et les nombres multiples sont conservés dans les quantités et les prix moyens.

- La méthode Q-U a été appliquée avec un mois de base fixe, à savoir le mois de décembre de chaque année comme dans l'IPC néerlandais. Le mois de base de 2014 est le mois de mars, car il s'agit du premier mois de la période choisie pour cette étude. Des intervalles de temps de 13 mois ont été utilisés (sauf bien sûr en 2014). Nous n'avons pas appliqué de mise à jour mais avons calculé les poids v_i et les indices de prix à l'aide des données complètes de tous les mois d'une année.

Le tableau 1 donne un exemple (pour des vêtements) de la façon dont les prix et les quantités des produits ont été calculés.

Le prix du produit est calculé à partir des données de caisse comme valeur unitaire, c'est-à-dire qu'il s'agit du rapport entre la somme des dépenses des six articles et la somme des quantités. Les dépenses et les quantités des articles retournés sont exclues, de sorte que ces valeurs sont additionnées avec

Tableau 1
Calcul des prix et des quantités de produits

Article	N° 1	N° 2	N° 3	N° 4	N° 5	N° 6	Produit
A – Données de caisse							
Dépense nette	0	118	13 201	2 711	25 108	13 009	-
Dépenses retournées	75	3 377	7 174	2 257	7 481	15 004	-
Quantité nette	0	0	899	186	1 643	986	-
Quantités retournées	5	198	372	124	434	812	-
Dépense	75	3 495	20 375	4 968	32 589	28 013	89 515
Quantité	5	198	1 271	310	2 077	1 798	5 659
Prix	14.95	17.65	16.03	16.03	15.69	15.58	15.82
B – Données moissonnées sur le Web							
Nombre de prix moissonnés	5	22	31	31	31	29	149
Somme des prix moissonnés	74.75	392.21	523.22	626.02	523.22	557.57	2 696.99
Prix	14.95	17.83	16.88	20.19	16.88	19.23	18.10

Note : pour les données de caisse : dépenses et quantités, tant pour les valeurs nettes que pour les retours de tee-shirts à manches courtes de la même marque. Les six articles ont des codes d'article différents (N° 1 à 6), qui sont combinés dans le même produit en fonction de caractéristiques communes. Les dépenses totales, la quantité totale et le prix (valeur unitaire en euros) du produit sont également indiqués. Les valeurs sont tirées des données de caisse du magasin en ligne et concernent un mois. Pour les données moissonnées sur le Web : nombres de prix moissonnés et somme de ces prix pour les mêmes articles et le même mois que pour les données de caisse. Ces valeurs sont également indiquées pour le produit ; elles correspondent à la somme des six articles.
Source : données de caisse et données moissonnées sur le Web.

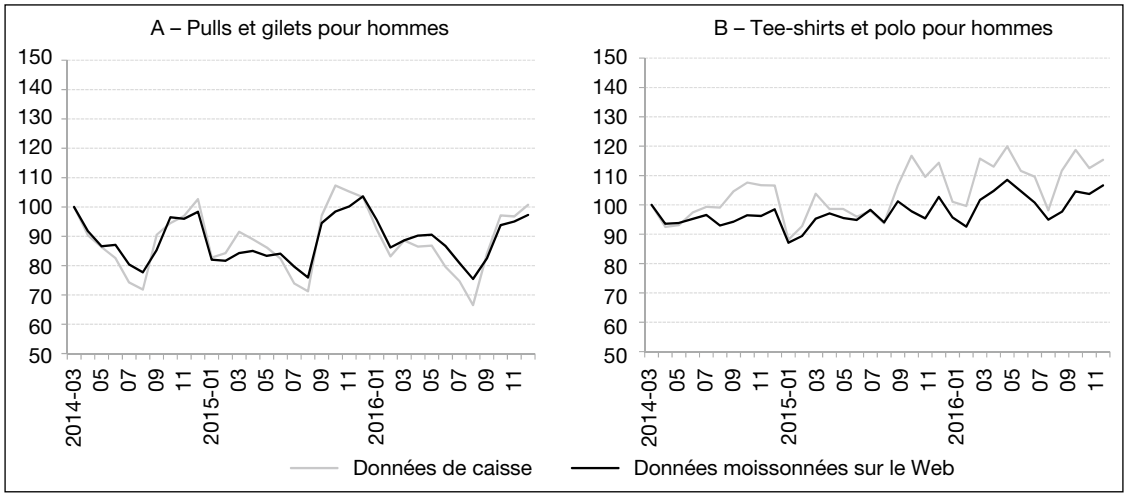
les quantités et les dépenses nettes. Pour le prix à partir des données moissonnées sur le Web (cf. tableau 1, B), il s'agit du rapport entre la somme des prix moissonnés durant les jours du mois et le nombre total des prix des articles moissonnés, additionnés pour les six articles (cf. dernière rangée du tableau 1, B).

Indices de prix

Les figures V et VI montrent les indices de prix calculés à partir de chaque source de données pour deux catégories de vêtements pour hommes et pour femmes. Les indices de prix calculés à partir des données moissonnées sur le Web suivent plus ou moins ceux calculés à partir des données de caisse, même les pics et creux des indices de données de caisse. La forte corrélation des prix comme des quantités entre données de caisse et données du Web se retrouve dans la comparaison des indices de prix. La correspondance étroite entre les indices de prix basés sur les deux jeux de données se retrouve également dans la totalité des 16 catégories de produits (voir en annexe les indices de prix de toutes les catégories de produits).

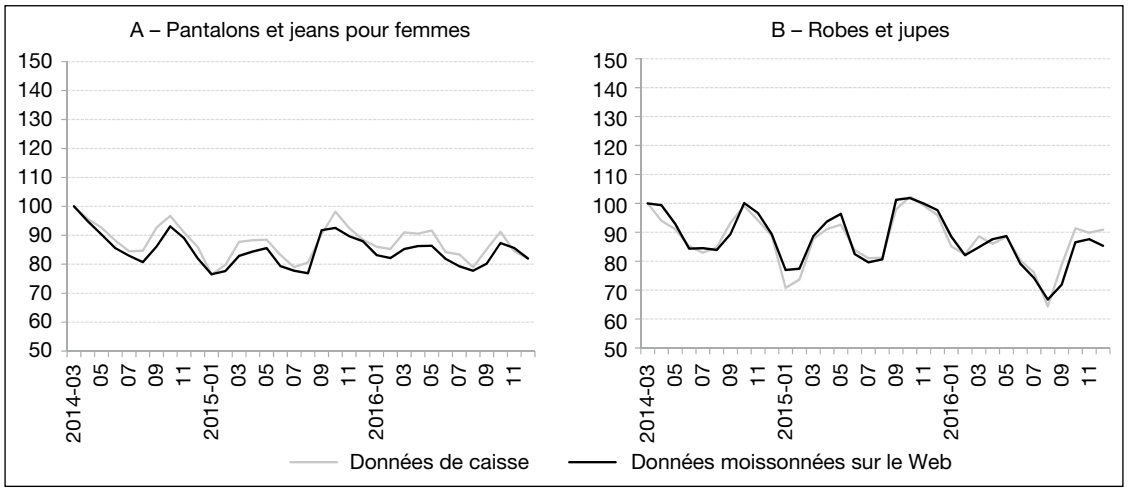
Les indices de prix des catégories de produits ont été combinés en appliquant l'habituelle méthode de Laspeyres, pour les vêtements pour hommes

Figure V
Indices Q-U pour deux catégories de vêtements pour hommes, pour les données de caisse et les données moissonnées sur le Web (mars 2014 = 100)



Source : données de caisse et données moissonnées sur le Web pour des vêtements.

Figure VI
Indices Q-U pour deux catégories de vêtements pour femmes, pour les données de caisse et les données moissonnées sur le Web (mars 2014 = 100)



Source : données de caisse et données moissonnées sur le Web pour des vêtements.

et les vêtements pour femmes classés en fonction de la nomenclature COICOP, et sont indiqués figure VII. S'agissant des données de caisse, on utilise des poids fixes annuels pour les catégories de produits. Les poids des catégories ont été fixés à un niveau égal à la part annuelle des dépenses de la catégorie concernée pour l'année précédente, excepté pour l'année 2014, première de la série, pour laquelle nous avons pris la part annuelle des dépenses de 2014.

Avec les données du Web, nous avons remplacé les dépenses par le prix moyen multiplié par le nombre de prix de produits moissonnés, additionné pour tous les produits au sein d'une catégorie donnée durant une année. Les différences entre les indices basés sur des données de caisse et ceux basés sur des données du Web sont très faibles pour les deux catégories COICOP. Entre les indices en glissement annuel, elles sont en moyenne seulement de 0.3 point de pourcentage, pour les deux catégories COICOP.

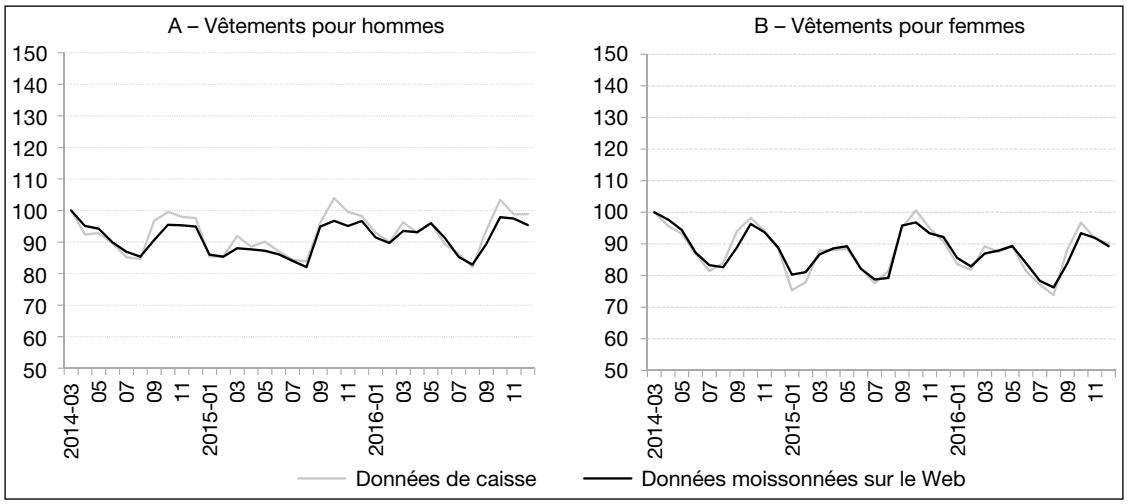
Analyse de sensibilité

Les résultats indiquent que, en utilisant le nombre de prix de produits moissonnés sur le Web au lieu du nombre de produits vendus, on obtient des indices de prix fiables. Cette conclusion correspond aux résultats de l'analyse des données fournie en première partie de cet article. Pour aller plus loin, nous avons cherché à déterminer si le fait de remplacer le nombre de prix de produits moissonnés par un nombre excluant toute corrélation avec le nombre de produits vendus était susceptible d'affecter les indices de prix. Nous avons remplacé le nombre de prix moissonnés par 0 ou 1, 0 indiquant que l'outil de moissonnage n'avait trouvé aucun prix pour un produit donné durant un mois donné et 1 indiquant que des prix avaient été trouvés mais que les nombres exacts avaient été ignorés. L'impact de ce changement sur les indices de prix est illustré figure VIII. Les résultats sont montrés uniquement au niveau de la nomenclature COICOP.

Le fait de remplacer le nombre de prix de produits moissonnés par 0 ou par 1 a un impact considérable sur les indices basés sur des données moissonnées sur le Web, ce qui est clairement visible au niveau de la nomenclature COICOP. Les résultats ne sont pas présentés pour les 16 catégories de produits, mais des différences du même ordre ont été identifiées pour 13 d'entre elles. Chacun de ces cas montre une tendance à la baisse de l'indice (comme dans la figure VIII).

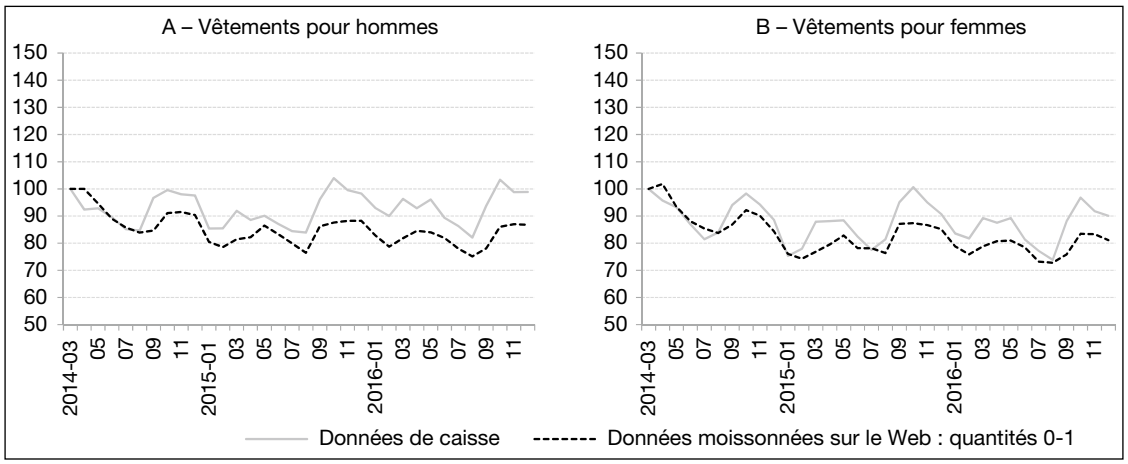
Les différences entre les indices d'une année à l'autre sont beaucoup plus importantes que celles obtenues avec les nombres initiaux de prix moissonnés. La différence moyenne avec les indices basés sur des données de caisse passe à près de 5 points de pourcentage pour les vêtements pour hommes et à près de 4 points de pourcentage pour les vêtements pour femmes. Ces résultats suggèrent que les nombres initiaux de prix moissonnés sur le Web devraient être utilisés pour calculer des indices de prix à partir de données moissonnées sur le Web. La manipulation de ces nombres, comme par exemple le retrait des prix dupliqués, est à éviter.

Figure VII
Indices de prix des vêtements pour hommes et pour femmes (mars 2014=100)



Source : données de caisse et données moissonnées sur le Web pour des vêtements.

Figure VIII
Indices de prix pour des vêtements pour hommes et pour femmes, avec le nombre de prix de produits moissonnés sur le Web remplacé par une valeur binaire (mars 2014 = 100)



Source : données de caisse et données moissonnées sur le Web pour des vêtements.

* *
*

À notre connaissance, l'étude ici présentée est la première à comparer des indices de prix calculés à partir de données de caisse et à partir de données moissonnées sur le Web. Cette comparaison a été possible parce que les deux sources de données sont disponibles pour le même détaillant. Ces premiers résultats semblent très prometteurs car les indices basés sur des données moissonnées sont d'une exactitude remarquable, surtout au niveau de la nomenclature COICOP. C'est d'autant plus utile que le recours au moissonnage du Web est de plus en plus envisagé pour les statistiques publiques. Les données de caisse restent l'option privilégiée car elles contiennent des données de transaction, mais tous les INS n'y ont pas nécessairement accès.

Ces résultats positifs et intéressants mettent encore plus l'accent sur la nécessité de comprendre pourquoi les indices de prix calculés uniquement à partir de données moissonnées sont si proches de ceux obtenus à partir de données de caisse. À ce stade, nous ne pouvons que spéculer sur les raisons possibles. Cela pourrait par exemple découler du fait que le détaillant est un magasin en ligne dépourvu de point de vente physique et qu'il pourrait donc être plus disposé à promouvoir les articles les plus vendus de son assortiment. Ces articles peuvent être agencés de façon à être rapidement identifiés par le consommateur sur le site Web, en les plaçant dans différents groupes principaux ou catégories. Par exemple, le même article peut être placé dans le groupe principal « En promotion » et dans l'un des autres groupes principaux classiques. Cela expliquerait, du moins en partie, la forte corrélation entre le nombre de produits vendus et le prix des produits moissonnés. Pour vérifier si un détaillant est plus susceptible de promouvoir les articles les plus vendus, il faudrait prendre contact avec lui afin qu'il explique la stratégie d'agencement de son site Web.

De façon plus générale, nous pouvons tirer plusieurs enseignements de cette étude.

- La méthode d'échantillonnage des prix sur un site Web est indéniablement importante. Cette étude montre, tout au moins pour le détaillant étudié dans cet article, que le moissonnage d'un site Web entier favorise l'exactitude des indices de prix calculés à partir de données du Web. Certes, il faut du temps pour moissonner un site Web entier, mais les instituts de statistique pourraient envisager de procéder à l'échantillonnage seulement certains jours et non quotidiennement.

- Le site Web traité dans cette étude a été moissonné par navigation, grâce à un outil de moissonnage de première génération développé par CBS. Cette technique prend elle aussi du temps, ce qui explique en partie la décision de moissonner la nuit. Les magasins en ligne utilisent une tarification dynamique. Les prix peuvent diminuer durant les heures d'ouverture, et tout prix manquant peut donc expliquer en partie les différences constatées entre les prix moissonnés et les prix tirés des données de caisse. Parallèlement, CBS a développé une deuxième génération

d'outils de moissonnage permettant d'extraire les prix et les métadonnées du code indiqué sur la page d'aperçu du produit. Cette technique est beaucoup plus rapide, ce qui permet de moissonner des sites Web de grande taille à différents moments de la journée. À l'avenir, cela permettra d'étudier l'impact de la tarification dynamique sur les indices de prix et de nous concentrer sur de nouvelles applications, comme par exemple la construction d'indices en temps réel. L'impact de la tarification dynamique sur les indices de prix est bien sûr impossible à quantifier ici. Toutefois, les différences constatées entre les indices de prix basés sur des données de caisse et ceux basés sur des données du Web sont faibles, ce qui suggère que cet impact est peu important dans le cas étudié ici.

- Cette étude suggère également d'utiliser le nombre initial de prix moissonnés pour calculer un indice de prix avec des données moissonnées sur le Web. La déduplication des prix est à éviter. Les résultats indiquent que les indices basés sur des données moissonnées perdent de leur exactitude lorsque l'on supprime les prix multiples (cf. figure VIII) : la différence avec les indices d'une année sur l'autre basés sur des données de caisse augmente alors à cinq points de pourcentage par an. En outre, tous les indices affichant un écart montrent une dérive à la baisse. Cela dit, nous admettons que la suppression des prix multiples a été assez extrême, dans la mesure où nous n'avons laissé qu'une seule observation par produit par mois. Néanmoins, les résultats indiquent que le nombre initial de prix moissonnés devrait être géré avec prudence.

- Il est toujours intéressant de demander aux détaillants de fournir des données sur les dépenses, même s'ils ne peuvent – ou ne veulent – pas fournir des données de caisse complètes.

Malgré tout, nous devons formuler des conclusions prudentes. En effet, les données moissonnées sur le Web ne sont pas des données de transaction et les résultats de cette étude ne concernent qu'un seul détaillant. L'analyse développée ici pourrait donc être complétée dans deux directions.

Tout d'abord, elle pourrait être élargie à d'autres magasins en ligne dont le site Web présente une structure semblable à celle analysée dans cet article, c'est-à-dire où les articles à prix réduit sont mis en avant plus souvent que les autres articles et où les articles les plus populaires sont plus faciles à trouver. L'unité IPC de CBS développe actuellement des outils de moissonnage du Web destinés aux détaillants vendant des appareils électroniques dont les données de caisse sont disponibles. Cela serait un cas d'étude intéressant, d'autant plus que ces détaillants ont des boutiques physiques. Promeuvent-ils les articles les plus vendus plus souvent que les articles moins populaires sur leur site Web ? Ou bien appliquent-ils une stratégie différente, par exemple en faisant la publicité des nouveaux articles ?

Le moissonnage du Web est un moyen utile pour compléter les informations collectées sur les articles dans les données de caisse, qui peuvent être limitées. En combinant les deux sources de données, on peut profiter de tous leurs avantages : des données de transaction fournies par les données de caisse et des informations supplémentaires sur les caractéristiques des articles fournies par les données moissonnées. En principe, ces conditions sont idéales pour appliquer et tester des méthodes visant à sélectionner les caractéristiques des articles et à définir des produits homogènes, permettant donc de traiter les relances. Toutefois, lorsque l'on utilise des métadonnées moissonnées pour compléter celles tirées des données de caisse fournies par des magasins physiques, il peut s'avérer impossible de compléter tous les GTIN inclus dans les données de caisse par des données moissonnées sur le Web. Les assortiments des magasins physiques et en ligne peuvent être différents, par exemple si les détaillants ne veulent inclure sur leur site Web qu'une partie des articles proposés dans les magasins physiques.

Pour finir, nous sommes conscients du fait que les études comparatives telles que celle présentée ici peuvent être difficiles à répliquer, car il est assez rare de disposer à la fois les données de caisse et les données moissonnées sur le Web pour un même détaillant. C'est encore plus difficile pour les INS rencontrant des problèmes dans l'acquisition des données de caisse. Pour cette raison, nous encourageons les INS ayant la chance de posséder des données de caisse à investir dans la recherche statistique sur données de caisse. Est-il possible, grâce à des analyses et tests statistiques, de caractériser les données de caisse ? Est-il possible d'identifier des tendances spécifiques, par exemple sur les corrélations entre les prix et les quantités au fil du temps ? En appliquant les mêmes analyses aux données moissonnées sur le Web, on pourrait évaluer leurs similarités avec les données de caisse et leur pertinence pour le calcul des indices de prix. Nous suggérons donc de donner une plus grande importance à l'analyse des séries temporelles et à d'autres analyses des données de caisse. □

BIBLIOGRAPHIE

Auer, L. von (2014). The Generalized Unit Value Index Family. *Review of Income and Wealth*, 60, 843–861. https://doi.org/10.1111/roiw.12042

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, M., Payne, C., Smith, T., Winton, J. & Woods, A. (2016). Research indices using web scraped data: May 2016 update. Office for National Statistics, internal report, 23 May 2016. https://www.ons.gov.uk/releases/researchindicesusingwebscrapedpricedatamay2016update

Cavallo, A. F. (2016). Are online and offline prices similar? Evidence from large multi-channel retailers. NBER, *Working Paper* N° 22142. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_MIT_are_online_and_offline_prices_similar.pdf

Chessa, A. G. (2016a). A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1/2016, 49–69. https://ec.europa.eu/eurostat/cros/content/new-methodology-processing-scanner-data-dutch-cpi-antonio-g-chessa_en

Chessa, A. G. (2016b). Comparisons of the QU-method with other index methods for scanner data. Paper prepared for the first meeting on multilateral methods organised by Eurostat, Luxembourg, 7-8 December 2016. Statistics Netherlands, Internal paper.

Chessa, A. G. (2017a). Comparisons of QU-GK indices for different lengths of the time window and updating methods. Paper prepared for the second meeting on multilateral methods organised by Eurostat, Luxembourg, 14-15 March 2017. Statistics Netherlands, Internal paper.

Chessa, A. G. (2017b). The QU-method: A new methodology for processing scanner data. *Statistics Canada International Symposium Series : Proceedings*. https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X201700014752

Chessa, A. G., Verburg, J. & Willenborg, L. (2017). A comparison of price index methods for scanner data. Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, Eltville am Rhein, Germany, 10-12 May 2017. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/A comparison of price index methods for scanner data -Antonio Chessa, Johan Verburg, Leon Willenborg -Paper.pdf

Daas, P. J. H. & van Nederpelt, P. W. M. (2010). Application of the object oriented quality management model to secondary data sources. Statistics Netherlands, the Hague/Heerlen, The Netherlands, *Discussion paper* N° 10012.

Daas, P. J. H. & Ossen, S. J. L. (2010). In search of the composition of data quality in statistics and other research areas. Statistics Netherlands, *Discussion paper*.

Diewert, W. E. & Fox, K. J. (2017). Substitution bias in multilateral methods for CPI construction using scanner data. Vancouver School of Economics, The University of British Columbia, *Discussion paper* N° 17-02. https://irs.princeton.edu/sites/irs/files/Diewert and Fox Substitution Bias and MultilateralMethodsForCPI_DP17-02_March23.pdf

Eurostat (2017). *Practical Guide for Processing Supermarket Scanner Data*. September 2017. https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

Geary, R. C. (1958). A note on the comparison of exchange rates and purchasing power between countries. *Journal of the Royal Statistical Society A*, 121, 97–99. https://doi.org/10.2307/2342991

Griffioen, A. R. & ten Bosch, O. (2016). On the use of internet data for the Dutch CPI. Paper presented at the *UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices*, Geneva, Switzerland, 2-4 May 2016. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_Netherlands_on_the_use_of_internet_data_for_the_Dutch_CPI.pdf

Griffioen, A. R., ten Bosch, O. & Hoogteijling, E. H. J. (2016). Challenges and solutions to the use of internet data in the Dutch CPI. Paper presented at the *UNECE Workshop on Statistical Data Collection,* The Hague, The Netherlands, 3-5 October 2016. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2016/mtg1/WP2-3_Netherlands_-_Griffioen_ap.pdf

de Haan, J., Willenborg, L. & Chessa, A. G. (2016). An overview of price index methods for scanner data. Paper presented at the *UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices*, Geneva, Switzerland, 2-4 May 2016. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_1_room_doc_Netherlands_an_overview_of_price_index_methods.pdf

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO Publications. https://doi.org/10.5089/9787509510148.069

Khamis, S. H. (1972). A new system of index numbers for national and international purposes. *Journal of the Royal Statistical Society A*, 135, 96–121. https://doi.org/10.2307/2345041

Krsinich, F. (2014). The FEWS Index: Fixed Effects with a Window Splice – Non-revisable quality-adjusted price indexes with no characteristic information. Paper presented at the *UNECE-ILO Meeting of the group of experts on consumer price indices*, Geneva, Switzerland, 26-28 May 2014. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/New_Zealand_-_FEWS.pdf

Lamboray, C. (2017). The Geary Khamis index and the Lehr index: how much do they differ? Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, Eltville am Rhein, Germany, 10-12 May 2017. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/The Geary Khamis index and the Lehr index how much do they differ - Claude Lamboray -Paper.pdf

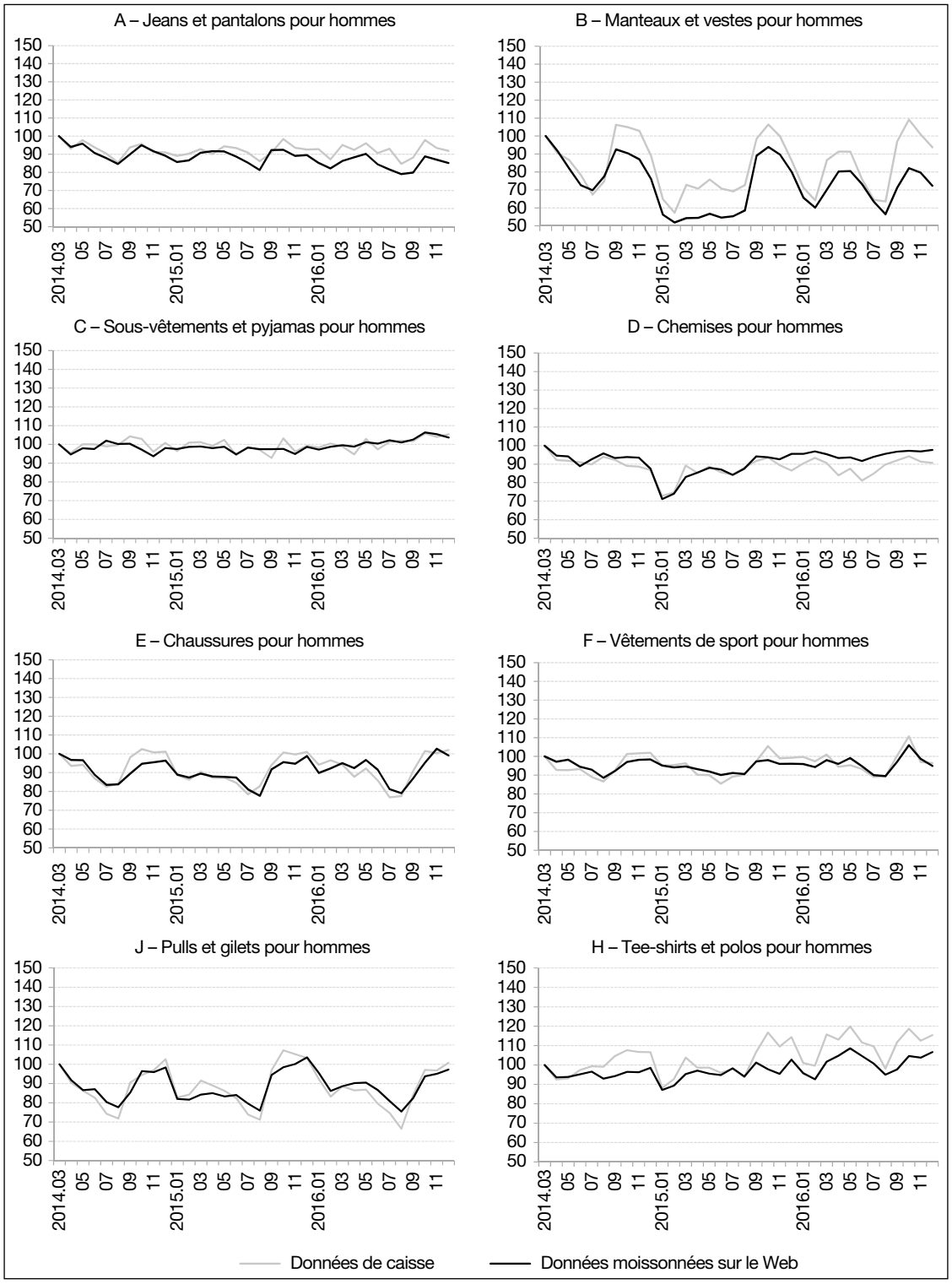
Maddison, A. & Rao, D. S. P. (1996). A generalized approach to international comparison of agricultural output and productivity. Groningen Growth and Development Centre, Research memorandum GD-27. https://www.rug.nl/research/portal/files/3258249/GD-27.pdf

Willenborg, L. (2017). Quantifying the dynamics of populations of articles. Statistics Netherlands, *Discussion Paper* N° 2017/10. https://www.cbs.nl/en-gb/background/2017/25/quantifying-the-dynamics-of-populations-of-articles

ANNEXE

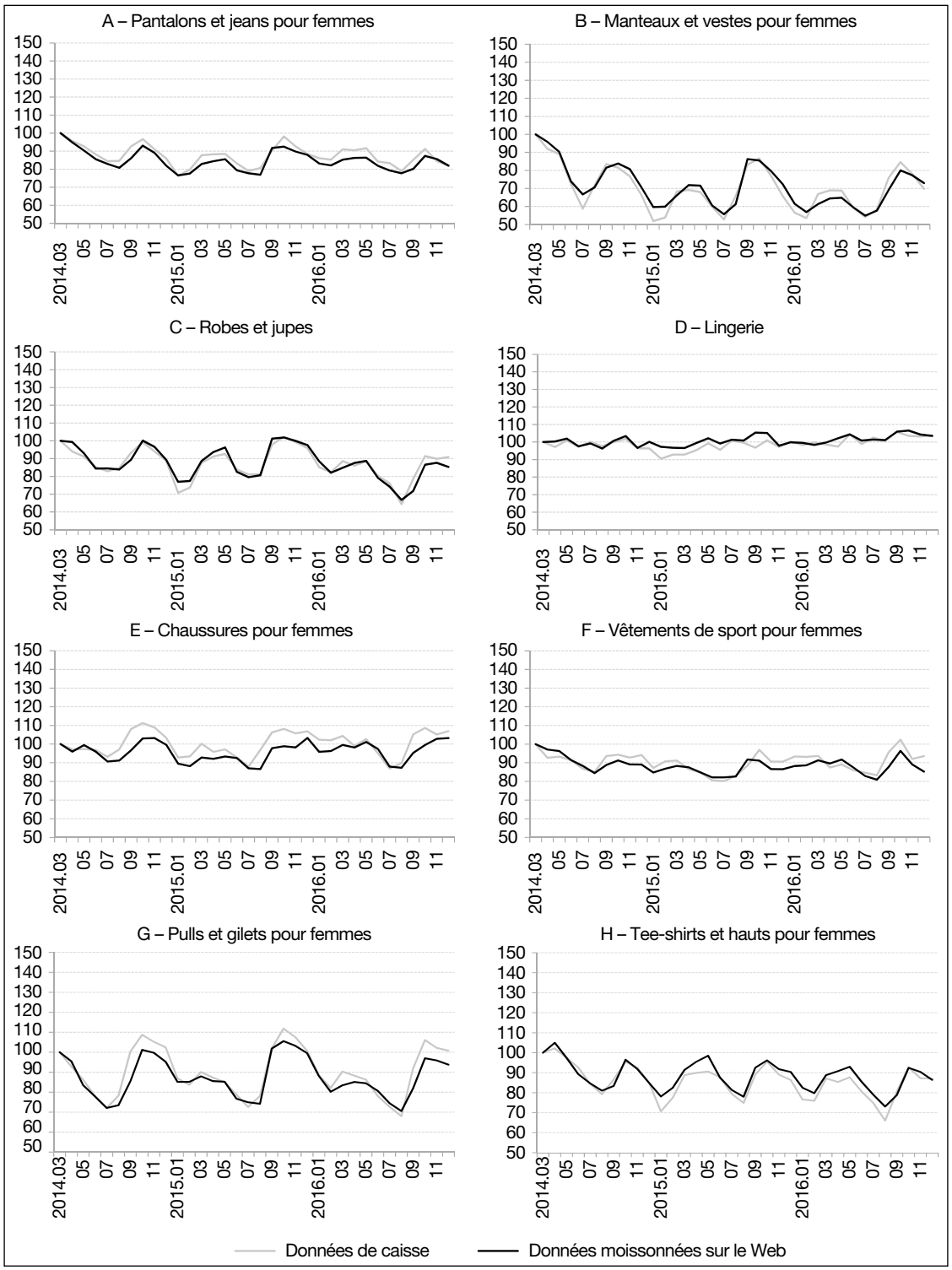
DONNÉES DE CAISSE ET DONNÉES MOISSONNÉES SUR LE WEB : INDICES DE PRIX POUR 16 CATÉGORIES DE PRODUITS DANS LE SECTEUR DES VÊTEMENTS POUR HOMMES ET POUR FEMMES

Figure A-I
Vêtements pour hommes (mars 2014 = 100)



Source : données de caisse et données moissonnées sur le Web pour des vêtements.

Figure A-II
Vêtements pour femmes (mars 2014 = 100)



Source : données de caisse et données moissonnées sur le Web pour des vêtements.

Écarts spatiaux de niveaux de prix entre régions et villes françaises avec des données de caisse

Spatial Differences in Price Levels between French Regions and Cities with Scanner Data

Isabelle Léonard*, Patrick Sillard*, Gaëtan Varlet* et Jean-Paul Zoyem*

Résumé – Cette étude s'appuie sur les données de caisse de la grande distribution transmises quotidiennement à l'Insee en 2013. Elle vise à calculer des indices mesurant des différences de niveau de prix à la consommation entre territoires métropolitains dans l'alimentaire vendu en supermarché. Un indice hédonique fondé sur la régression du prix du produit sur des indicatrices de codes-barres et de territoires est développé. Plusieurs évaluations sont déterminées sur différentes semaines, une semaine de données permettant déjà d'atteindre une précision considérable. La dispersion des niveaux de prix entre régions ou entre grandes agglomérations est limitée et, pour l'essentiel, robuste au choix de la semaine : les prix les plus élevés sont observés en région parisienne ainsi qu'en Corse, les écarts étant de l'ordre de quelques points de pourcentage. Le rapprochement de ces résultats nouveaux et des travaux réalisés par l'Insee entre 1970 et 2000 montre que les écarts de prix dans l'alimentaire entre territoires métropolitains sont essentiellement structurels et évoluent peu dans le temps.

Abstract *– This study is based on scanner data from large retailers sent daily to Insee in 2013. Its aim is to compute indices that measure differences in consumer price levels between different areas of metropolitan France, focusing specifically on food products sold in supermarkets. A hedonic index based on the regression of the product price on barcode and territory dummies is developed. Several assessments are carried out over different weeks, with one week of data already providing a great degree of accuracy. The dispersion of price levels between regions or large conurbations is limited and, for the most part, robust to the choice of week. The highest prices are found in the Paris region and Corsica, with a magnitude of differences in the order of a few percentage points. A comparison of the new findings with research conducted by Insee between 1970 and 2000 shows that differences in food prices across areas of metropolitan France are essentially structural and change little over time.*

Codes JEL / JEL Classification : E31, C8, D1
Mots-clés : niveaux de prix, comparaison spatiale, données de caisse
Keywords: price levels, spatial comparison, scanner data

Rappel :

Les jugements et opinions exprimés par les auteurs n'engagent qu'eux mêmes, et non les institutions auxquelles ils appartiennent, ni *a fortiori* l'Insee.

** Insee (isabelle.leonard@insee.fr ; patrick.sillard@insee.fr ; gaetan.varlet@insee.fr ; jean-paul.zoyem@insee.fr)*

Reçu le 18 septembre 2017, accepté après révisions le 17 janvier 2019

Pour citer cet article : Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2019). Spatial Differences in Price Levels between French Regions and Cities with Scanner Data. *Economie et Statistique / Economics and Statistics*, 509, 69–82. https://doi.org/10.24187/ecostat.2019.509.1983

Le système d'observation des prix à la consommation mis en place par la statistique publique française vise essentiellement à déterminer les écarts temporels de prix, c'est-à-dire l'inflation. L'indice des prix à la consommation (IPC) est une mesure de cette grandeur. Pour cela, mois après mois, les enquêteurs de l'Insee retournent dans les mêmes points de vente observer le prix des mêmes produits, et c'est à partir des évolutions élémentaires de prix observées au niveau de chaque produit suivi dans l'IPC qu'est calculée l'évolution moyenne d'ensemble. L'intuition suggère que les prix collectés pour les besoins de l'IPC pourraient aussi permettre de déterminer des écarts moyens de niveaux de prix entre différents territoires d'intérêt. Pourtant ce n'est pas en général le cas. En effet, l'enjeu, pour mesurer une évolution moyenne des prix, est d'assurer qu'entre deux périodes comparées, on compare effectivement les mêmes produits. Dans le même esprit, comparer des niveaux de prix entre territoires suppose d'observer les prix de produits identiques sur les territoires dont les niveaux de prix sont comparés. Comme ce dernier enjeu, spécifique à la comparaison de niveaux territoriaux, n'en est pas un pour l'IPC, l'identification des produits réalisée pour l'IPC n'est en général pas suffisamment riche pour assurer que deux produits observés dans deux points de vente différents sont identiques. Par ailleurs, l'échantillon de produits suivis dans l'IPC est obtenu par sondage et optimisé pour atteindre une précision satisfaisante de la mesure d'inflation au plan national. Descendre plus finement en localisation géographique se heurte mécaniquement à la faiblesse du nombre de relevés effectués sur des territoires d'extension limitée. Au final, même si les produits étaient mieux identifiés dans l'IPC, on ne pourrait que difficilement effectuer des comparaisons satisfaisantes de niveaux de prix entre territoires.

À l'opposé, les données de caisse ne comportent pas certaines des limitations évoquées au sujet des données de l'IPC pour déterminer des écarts de niveaux de prix spatiaux : 1) le code-barres (appelé aussi EAN pour *European Article Number*) est un identifiant univoque du produit[1] ; 2) les données de caisse couvrent l'exhaustivité des transactions relatives aux produits de l'alimentaire industriel[2] hors produits frais – i.e. les fruits, légumes, crustacés ainsi qu'une partie des poissons et des viandes –, aux boissons alcoolisées ou non et à une partie des biens manufacturés vendus dans les hypermarchés et supermarchés de France métropolitaine. La première propriété précédente assure que la comparaison de prix d'un même code-barres vendu dans deux magasins différents conduit à comparer *de facto* le même produit. La seconde propriété assure que les échantillons disponibles sont suffisamment vastes pour permettre une comparaison à des niveaux de détail fins.

L'Insee a engagé une expérience pilote destinée à intégrer progressivement les données de caisse dans la détermination de l'IPC. Pour cela, l'Insee reçoit quotidiennement depuis la fin de l'année 2012 les données de caisse de plusieurs groupes de la grande distribution. Les groupes participant à l'expérience pilote représentent environ 30 % du champ potentiel, c'est-à-dire celui qui correspondrait aux transactions quotidiennes de l'ensemble des enseignes de la grande distribution présentes sur le territoire métropolitain. Les données de caisse regroupent, pour chaque magasin, la liste des transactions du jour, c'est-à-dire la liste des codes-barres vendus, ainsi que les quantités écoulées et les prix de vente[3] correspondants.

Un des avantages essentiels des données de caisse est la richesse d'information. L'extraordinaire volume de données permet d'envisager de fournir un niveau de détail d'information sur la connaissance des niveaux de prix inatteignable avec le dispositif habituel de collecte. En outre, ces données comportent simultanément l'information de prix et celle des quantités de produits vendus, ce qui constitue une matière inédite pour la statistique de prix puisque celle-ci se fonde, habituellement, sur la seule connaissance des prix de détail. Si les premières applications portent naturellement sur la détermination de l'inflation dans les pays (voir par exemple Reinsdorf, 1999 ; de Haan & van der Grient, 2011), d'autres applications statistiques sont possibles. La comparaison de niveaux de prix entre pays reste compliquée car les produits élémentaires, le système de codage des produits ou simplement les systèmes d'information des enseignes ne sont généralement pas suffisamment homogènes pour permettre un rapprochement en masse des EAN. En revanche, pour un pays donné, lorsque le système d'information des données de caisse fournit en même temps des informations détaillées selon les lieux d'achat, il devient possible d'utiliser les données de caisse pour déterminer des écarts de niveau de prix entre zones géographiques. C'est cette

1. C'est-à-dire que deux produits différents (vus comme tels par le consommateur) ont nécessairement deux EAN différents. En revanche, deux EAN différents peuvent désigner le même produit.
2. Sauf mention contraire, on entend, dans cet article, le champ de l'alimentaire industriel comme étant celui des produits d'alimentation, hors produits frais (i.e. fruits et légumes frais, crustacés ainsi qu'une partie des poissons et des viandes), et des boissons alcoolisées ou non, vendus en supermarchés (voir la section sur les données pour plus de détails).
3. Parfois le chiffre d'affaires correspondant, plutôt que le prix.

question qui est examinée ici, pour le champ de l'alimentaire industriel, à l'aide d'un jeu de données de caisse dont l'Insee a pu disposer sur l'année 2013.

La comparaison spatiale de niveaux de prix est une pratique courante dans de nombreux pays, généralement coordonnée par des institutions internationales. S'agissant d'indices bilatéraux, l'opération consiste à définir des classes de produits homologues entre les pays, à déterminer une structure de consommation en dépense pour le couple de pays considérés, à identifier des produits représentatifs des consommations nationales et comparables dans leur usage entre les deux pays, puis à calculer un indice bilatéral caractérisant la différence de niveaux de prix entre pays. Une des grandes difficultés de ce genre d'opération est de déterminer des classes de produits réellement homologues, c'est-à-dire qui correspondent à un « usage » équivalent dans les différents pays comparés. En effet, faute d'être capable d'identifier des produits identiques – lesquels n'existent d'ailleurs pas toujours, en particulier lorsque les pays sont assez différents dans leurs cultures et niveaux de vie –, les institutions qui coordonnent ces comparaisons fondent la mesure des différences de prix sur le rapprochement de produits dont les caractéristiques sont aussi proches que possible. Si cette approche donne une bonne approximation des écarts de niveaux de prix, fondée sur un compromis entre définition des produits et comparabilité, elle reste cependant discutable du fait de ce compromis. Les limites de ces comparaisons dites en « parité de pouvoir d'achat » sont bien connues et détaillées dans la littérature (voir par exemple Deaton & Heston, 2010). Il faut retenir de cette littérature que les discussions portent sur deux points, de nature différente mais qui, tous deux, sont d'importance limitée pour l'exercice de comparaison conduit ici sur des données de caisse et pour comparer des régions françaises entre elles. Le premier point de débat concerne l'exercice de rapprochement des produits qui devient potentiellement impossible lorsque les zones comparées sont très différentes ; dans le cas qui nous intéresse ici, les zones comparées – des régions de France métropolitaine ou des agglomérations – sont très homogènes dans leurs usages de consommation. Le second point porte sur la méthode de calcul des indices d'écart de niveaux. En pratique, les méthodes utilisées conduisent à des indices qui diffèrent d'autant moins que les prix et structure de consommation sont proches entre les zones comparées.

Potentiellement plus critique est la question du champ de la comparaison. Par construction, les résultats présentés dans cet article portent sur le champ sur lequel on dispose de données de caisse. D'une part, il s'agit en l'occurrence du champ des produits de l'alimentaire (hors produits frais) et des boissons alcoolisées ou non, vendus en grandes surfaces (i.e. alimentaire industriel). Les achats alimentaires réalisés dans d'autres types de points de vente ne sont pas inclus et les résultats obtenus ne sont donc pas représentatifs de l'ensemble de la consommation alimentaire. D'autre part, en 2013, l'Insee ne disposait des données de caisse que de quelques enseignes de la grande distribution. Les ventes correspondantes représentaient environ 30 % des ventes de la grande distribution, dans l'alimentaire industriel, en métropole. Par conséquent, il est possible que les comparaisons régionales de niveau de prix examinées dans cet article soient affectées d'un biais lié à cette sélection d'enseignes. La section consacrée à la présentation des données examine plus en détail ces questions de couverture et montre, en particulier, que la structure de consommation obtenue à partir du champ restreint est conforme à la répartition géographique de la population française. L'impact éventuel de la politique géographique de prix des enseignes sélectionnées dans l'échantillon est plus compliqué à cerner : si la politique est spécifique à l'enseigne et que, dans le même temps, le poids de cette enseigne dans le territoire comparé diffère entre l'échantillon de l'Insee et la situation générale toutes enseignes réunies, alors l'indice du territoire estimé sur fondement de l'échantillon particulier sera différent de celui obtenu toutes enseignes confondues. *A priori*, les effets de la concurrence locale tendent toutefois à forcer à l'harmonisation des structures de prix entre enseignes et dans l'espace. Par conséquent, des estimations fondées sur un sous-échantillon couvrant 30 % de la population d'ensemble devraient, dans ce contexte, permettre de tirer des enseignements de portée relativement générale.

La suite de l'article est organisée ainsi : une première section présente les résultats d'autres exercices de mesure des écarts de prix entre régions métropolitaines et entre grandes agglomérations réalisées par l'Insee depuis 1971. Les résultats nouveaux obtenus à partir des données de caisse dans cette étude sont ainsi mis en perspective de résultats comparables et plus anciens. Des statistiques descriptives sont données dans la deuxième section, puis la section suivante présente la modélisation retenue pour l'analyse des données. Une dernière section présente les différents résultats obtenus et une analyse de robustesse, qui reprend les différents axes de discussion évoqués plus haut.

Quelques expériences passées de comparaisons spatiales au niveau du territoire métropolitain

Les travaux de comparaison de niveaux de prix entre régions métropolitaines sont très anciens puisque figurent, dans les publications de la Statistique Générale de la France (SGF), à la fin du 19e siècle et au début du 20e, des tableaux comparatifs de prix de détail moyens par denrées observés dans différentes villes de France. Il faut toutefois attendre des années plus récentes pour disposer de comparaisons qui couvrent un champ significatif de la consommation et qui sont fondées sur un nombre important de produits. Techniquement, ces travaux[4] consistent, s'agissant des comparaisons de niveaux de prix métropolitains, à calculer un ratio de prix moyen entre le territoire considéré et la France entière, pour des produits représentatifs de la consommation d'une variété de produits donnée, puis à agréger les différences ainsi mesurées au niveau des variétés de produits en une moyenne pondérée nationale. La pondération appliquée pour cette moyenne correspond à la structure de consommation nationale, sans particularités locales, au motif que les structures locales de consommations sont très peu différentes de la structure nationale (Mineau, 1987 ; Anxionnaz & Mothe, 2000). Parmi les travaux plus récents que ceux de la SGF, on peut citer Piccard (1972) et Baraille (1978) qui traitent des différences de niveaux entre villes métropolitaines. Les résultats de ces deux études sont repris dans le tableau 1. Ces deux études concluent, dans le même sens, que dans le champ de l'alimentaire et des boissons, l'agglomération parisienne et la Corse sont les lieux où les prix de l'alimentaire et des boissons (alcoolisées ou non) sont les plus élevés de métropole. Par ailleurs, elles mettent en évidence une dispersion somme toute peu importante, comprise dans une fourchette un peu inférieure à 10 points de pourcentage[5].

Ces travaux ont été complétés par Baraille & Bobin (1981) en utilisant une analyse par type de territoire basée sur une nouvelle enquête réalisée en 1981. Ce type d'analyse faisait écho à des résultats du même type obtenus par Piccard (1972).

Plus récemment, Mineau (1987) donne la ventilation par grande agglomération des écarts de niveaux de prix dans l'alimentaire et les boissons pour l'année 1985 ; la Division prix de détail de l'Insee (1990) en fait de même pour l'année 1989. Ces deux groupes de résultats témoignent d'une stabilité des écarts en niveaux de prix entre les différentes agglomérations que l'on peut constater au tableau 1. Naturellement, les deux années étudiées (ici 1985 et 1989) sont proches mais le constat est similaire avec l'année 1977 pourtant plus éloignée (tableau 1). À nouveau, dans ces travaux, on constate que les prix dans l'alimentaire et des boissons sont plus élevés en Corse que partout ailleurs. Puis, vient l'agglomération parisienne dans laquelle les prix à la consommation sont de 2 à 3 % plus élevés que dans les autres villes de province.

L'étude sur 1995 menée par Guglielmetti (1996) établit quant à elle que l'écart de niveau moyen des prix de l'alimentaire et des boissons (alcoolisées ou non, y. c. tabac) en Corse est sensiblement plus élevé qu'en 1989, puisqu'il atteint 8.5 % par rapport à Paris, l'écart entre Paris et Marseille s'étant maintenu sur la période.

Les derniers travaux menés, plus globaux, n'indiquent pas de modifications notables par rapport à ces constats. Fesseau *et al*. (2008) constatent que les prix de l'alimentaire et des boissons non alcoolisées sont environ 5.7 % plus élevés en Île-de-France qu'en province sur l'année 2006. Nicolaï (2010) établit, à partir de l'enquête de comparaison spatiale des niveaux de prix menée par l'Insee en 2010, que les niveaux moyens des prix de l'alimentaire et des boissons non alcoolisées sont environ 8.6 % plus élevés en Corse que sur le continent pris dans son ensemble. Enfin, la réédition de cette enquête en 2015 a permis de montrer que les prix de l'alimentaire[6] et des boissons non alcoolisées, cette année-là, étaient de 6.5 % plus élevés en région parisienne qu'en province et de 2.1 % plus élevés en Corse qu'en région parisienne (Clé *et al*., 2016). Ces derniers résultats, établis à partir de données collectées pour mesurer les écarts de niveaux de prix, confirment donc, dans le champ de l'alimentaire, la hiérarchie et les ordres de grandeurs établis précédemment.

Au final, ces différentes études dont le champ, la méthodologie et la nature diffèrent quelque peu, donnent des résultats globalement cohérents : les différences de niveaux de prix sont des

---

4. À l'exception notable des travaux les plus récents de comparaison spatiale de prix adossés à des enquêtes ad hoc (Nicolaï, 2010 ; Berthier et al.*, 2010 ; Clé* et al.*, 2016). Ces travaux s'appuient sur une approche inspirée des enquêtes européennes harmonisées de mesure des parités de pouvoir d'achat et sont fondés sur des indices de prix de Fisher, calés sur des structures de consommations spécifiques à chacun des territoires comparés. Cette approche se justifie pleinement lorsque les structures de consommation diffèrent significativement entre les territoires comparés, comme c'est le cas entre les DOM et la métropole par exemple. En revanche pour des comparaisons de régions métropolitaines, les différences de structures régionales sont très limitées et l'enjeu de leur prise en compte est secondaire.*

5. L'étude de Baraille (1978) mesurait un écart de 8 % entre les prix de l'alimentaire et des boissons dans l'agglomération où ils étaient les plus élevés (Ajaccio-Bastia) et les moins élevés (Angers).

6. Incluant également les produits frais.

Tableau 1
Écarts de prix moyens observés en France métropolitaine dans le champ de l'alimentaire et des boissons

Territoire	Indice, selon les résultats de :				
	Piccard, 1972 pour 1971	Baraille, 1978 pour 1977	Mineau, 1987 pour 1985	Insee, 1990 pour 1989	Guglielmetti, 1996 pour 1995
Agg. parisienne	100	100.0	100.0	100.0	100.0
Lyon	100	97.5	99.0	98.7	
Marseille	104	98.3	99.5	97.5	97.0
Bordeaux	100	94.1	96.7	96.6	
Rennes	97	93.8	92.8	94.4	
Reims		97.2	97.7	97.8	
Rouen		97.7	95.9	95.1	
Strasbourg		98.1	97.0	98.2	
Lille		97.6	95.3	95.7	
Orléans		95.7	96.2	95.7	
Limoges		97.4	96.7	97.1	
Ajaccio-Bastia		100.5	105.1	103.6	108.5
Clermont-Ferrand		99.0	100.9	98.5	
Toulouse		95.1	98.5	98.9	
Dijon		96.7	96.9	97.9	
Nantes		93.6	93.7	94.7	
Nancy		95.0	98.9	97.1	
Poitiers		94.2	92.5	92.2	
Montpellier		96.4	100.1	100.4	

Note : le niveau global des indices est fixé en référence à l'agglomération parisienne (recalculé par les auteurs à partir des données publiées pour référence à l'agglomération parisienne).

caractéristiques fortement structurelles qui évoluent donc assez peu au cours du temps ; les prix sont plus élevés en Corse, probablement en raison de la géographie insulaire qui limite la concurrence et renchérit les coûts de production notamment par les coûts de transport des produits élaborés sur le contient ; ils sont aussi plus élevés en région parisienne, cette fois sans doute en raison de coûts de commercialisation plus élevés (prix de l'immobilier commercial) et du pouvoir d'achat des résidents-consommateurs, en moyenne plus élevé qu'ailleurs.

Les données

Les données utilisées sont les données de caisse des enseignes signataires d'une convention avec l'Insee autorisant l'institut à accéder à des enregistrements quotidiens pour l'année 2013. Dans ces données, on ne retient pour cette étude que celles qui se rapportent à l'alimentaire industriel, c'est-à-dire aux produits d'alimentation et aux boissons, alcoolisées ou non[7], vendus en supermarchés. Ces données sont issues, en avril 2013, de 1 833 magasins. Ces magasins sont répartis dans 1 330 communes situées dans 707 agglomérations de métropole. La répartition du nombre de points de vente par grandes agglomérations[8] dans les études relatées plus haut est donnée au tableau 2.

La répartition par région est donnée dans le tableau 3. Notons qu'il s'agit, ici comme dans tout l'article, des régions administratives antérieures à la réforme (loi NOTRe) de 2015. Dans l'ensemble, la distribution du nombre de points de vente au niveau régional est relativement proche de celle de la démographie. Ainsi les points de vente de la base de données renvoient, par leur répartition géographique, une image assez fidèle de l'espace marchand français. Naturellement, dans la mesure où un nombre limité de groupes de la grande distribution ont transmis leurs données à l'Insee en 2013, des effets de grappes restent à craindre.

La structure de consommation, en termes de produits consommés, devrait *a priori* être voisine d'une région à l'autre. Afin d'examiner cette hypothèse,

7. Division de la nomenclature COICOP 01, hors produits frais (fruits et légumes frais, crustacés ainsi qu'une partie des poissons et des viandes) et groupe de la nomenclature COICOP 02.1.
8. Nomenclature des unités urbaines, version 2010. Cette nomenclature comprend environ 2 000 unités.

Tableau 2
Nombre de points de ventes par grande agglomération dans l'échantillon utilisé

Agglomération	Nombre de points de vente
Agglomération parisienne	352
Lyon	50
Marseille	31
Bordeaux	30
Rennes	10
Reims	8
Rouen	15
Strasbourg	19
Lille	26
Orléans	13
Limoges	4
Ajaccio-Bastia	4
Clermont-Ferrand	16
Toulouse	26
Dijon	4
Nantes	9
Nancy	5
Poitiers	2
Montpellier	12

Note : lorsque le nombre de points de vente est inférieur ou égal à 4 (Limoges, Dijon, Poitiers, Ajaccio-Bastia), l'indice de la ville ne figure pas dans la table de résultats (voir tableau 7).
Source : Insee, données de caisse 2013.

nous procédons au calcul de cette structure avec la base des données de caisse : le tableau 4 donne la répartition des chiffres d'affaires associés aux regroupements des produits selon la nomenclature internationale par fonctions de consommation (COICOP). Ces statistiques montrent, comme attendu, que les structures régionales, dans le champ de l'alimentaire industriel, s'écartent peu d'une structure moyenne métropolitaine portant sur le même champ. On note également que cette structure, propre aux achats réalisés en supermarchés, diffère sensiblement de la structure de consommation toutes formes de ventes confondues, essentiellement s'agissant des produits frais non industriels (fruits et légumes frais, crustacés, une partie des poissons et des viandes).

Ainsi construite, la base de données comprend, en moyenne, 16.4 millions d'observations par semaine, correspondant au croisement [point de vente × EAN], des prix moyens par codes-barres et des chiffres d'affaires. Le chiffre d'affaires total d'une semaine d'observation disponible dans la base de données est, en moyenne, d'environ 445 millions d'euros. Extrapolé sur une année (52 semaines) et rapporté à la

Tableau 3
Nombre de points de ventes par région dans l'échantillon utilisé

Région	Nombre de points de vente	Poids en nombre de PV (en %)	Poids démographique (en %)
Île-de-France	404	22.1	*18.8*
Rhône-Alpes	201	11.0	*10.0*
Nord-Pas-de-Calais	162	8.9	*6.4*
Provence-Alpes-Côte d'Azur	105	5.7	*7.8*
Centre	104	5.7	*4.0*
Aquitaine	94	5.1	*5.2*
Haute-Normandie	79	4.3	*2.9*
Picardie	73	4.0	*3.0*
Midi-Pyrénées	72	3.9	*4.6*
Bretagne	71	3.9	*5.1*
Auvergne	67	3.7	*2.1*
Languedoc-Roussillon	65	3.6	*4.2*
Basse-Normandie	58	3.2	*2.3*
Pays de la Loire	51	2.8	*5.7*
Lorraine	44	2.4	*3.7*
Alsace	44	2.4	*2.9*
Champagne-Ardenne	36	2.0	*2.1*
Bourgogne	33	1.8	*2.6*
Limousin	25	1.4	*1.2*
Poitou-Charentes	21	1.1	*2.8*
Franche-Comté	15	0.8	*1.9*
Corse	5	0.3	*0.5*
Total	1 829	100	*100*

Lecture : dans la base de données utilisée, l'Île-de-France comprend 404 points de vente. Les 404 points de vente représentent 22.1 % des 1 829 points de vente que comprend la base de données. Pour rappel et comparaison, l'Île-de-France regroupe 18.8 % des habitants de France métropolitaine (Recensement de la population, 2012). Les chiffres donnés en italiques ne proviennent pas de la base de données de caisse.
Source : Insee, données de caisse 2013.

dépense[9] de consommation des ménages observée en 2012 dans l'alimentaire et les boissons alcoolisées ou non, ce chiffre d'affaires représente environ 15 % de la dépense de consommation des ménages du champ[10].

Modèle d'estimation

Une observation élémentaire correspond à un code-barres (EAN) vendu dans un magasin de l'échantillon durant la semaine considérée. Il y a donc une observation par [point de vente × EAN]. On suppose que les observations élémentaires ainsi définies sont repérées par un indice i d'un ensemble I. Ainsi p_i est le prix (valeur unitaire sur la semaine) de l'article repéré par son code-barres dans un des magasins de la base de données. On note ω_i le chiffre d'affaires associé à l'observation correspondante.

L'indice traduisant les écarts de niveaux de prix entre zones géographiques est calculé par une méthode hédonique (Triplett, 2006). Cette approche, fondée sur une modélisation économétrique des prix, diffère quelque peu des approches harmonisées appliquées pour la mesure des parités de pouvoir d'achat entre pays européens. Pour autant, elle figure au nombre des méthodes classiques (Deaton & Heston, 2010) et, dans le cas où les territoires comparés présentent des consommations proches (en prix et en structure – c'est le cas ici comme le montre le tableau 4), elle conduit à des mesures d'écarts de niveaux de prix voisines des méthodes alternatives.

Le modèle économétrique est conditionné par le code-barres et la zone géographique d'appartenance du produit i considéré. Grâce au conditionnement par le code-barres, le modèle

9. Source Comptabilité nationale, euros courants, soit 156 milliards d'euros.
10. Pour être précis, les écarts sur le champ tiennent aux produits alimentaires vendus dans les autres points de vente (de la Grande distribution pour des enseignes non intégrées dans l'étude car ne transmettant pas en 2013 leurs données à l'Insee, ainsi que d'autres types de magasins, ou des marchés) et aux produits frais.

Tableau 4
Structures régionales de consommation dans le champ de l'alimentaire industriel

Région	Code	01.1.1	01.1.2	01.1.3	01.1.4	01.1.5	01.1.6	01.1.7	01.1.8	01.1.9	01.2.1	01.2.2	02.1.1	02.1.2	02.1.3	Total
Île-de-France	11	13.3	10.2	5.4	19.4	3.0	1.1	6.0	7.7	2.6	3.4	11.0	5.4	9.1	2.6	100
Champagne-Ardenne	21	9.8	10.4	4.1	17.2	2.8	0.9	5.6	6.1	2.0	3.2	9.7	6.3	18.0	3.8	100
Picardie	22	10.7	11.6	4.6	18.3	3.4	0.8	5.9	6.1	2.2	3.3	11.0	9.1	9.2	3.7	100
Haute-Normandie	23	10.6	10.7	4.5	17.0	3.1	0.9	5.7	6.4	2.1	3.5	10.3	11.4	10.6	3.2	100
Centre	24	11.3	11.1	5.2	18.8	3.4	1.0	6.1	6.8	2.2	3.5	10.5	8.0	8.3	3.7	100
Basse-Normandie	25	11.1	9.7	4.5	17.5	3.3	1.0	5.8	7.0	2.0	3.8	9.0	9.5	12.6	3.3	100
Bourgogne	26	10.7	10.7	4.6	18.5	3.2	0.9	6.0	6.9	2.3	3.5	10.1	6.5	12.3	3.8	100
Nord-Pas-de-Calais	31	9.8	10.0	4.0	16.9	3.4	0.8	5.4	6.4	2.2	3.3	11.9	8.4	12.7	4.6	100
Lorraine	41	11.6	10.2	4.7	19.9	3.2	0.8	5.8	7.0	2.4	3.8	12.0	4.8	9.0	4.8	100
Alsace	42	11.7	9.3	4.6	19.8	3.5	1.0	5.6	7.2	2.9	3.8	13.5	4.4	7.6	5.2	100
Franche-Comté	43	11.1	10.3	5.1	17.9	3.3	1.0	6,1	7.2	2.3	3.9	10.3	5.5	11.6	4.6	100
Pays de la Loire	52	11.9	10.2	5.0	17.9	3.4	1.1	6,2	7.2	2.1	3.5	9.5	7.8	10.1	4.2	100
Bretagne	53	11.3	10.4	4.2	16.5	3.4	1.2	5.7	7.3	2.0	3.7	8.9	7.2	14.2	4.0	100
Poitou-Charentes	54	10.6	11.3	5.3	18.2	3.2	1.0	5.9	6.4	2.1	3.6	10.2	7.2	10.7	4.2	100
Aquitaine	72	11.6	10.6	5.7	18.7	3.3	1.1	6.4	7.0	2.2	4.0	10.1	5.5	9.5	4.3	100
Midi-Pyrénées	73	12.5	9.8	5.7	19.3	3.3	1.1	6.2	7.6	2.5	4.1	10.0	5.1	8.7	4.4	100
Limousin	74	10.5	9.7	4.8	17.7	3.4	1.1	5.7	6.9	2.1	3.8	9.4	7.5	13.3	4.2	100
Rhône-Alpes	82	12.4	9.7	5.4	18.9	3.3	1.0	5.7	7.8	2.6	3.5	10.3	5.3	9.9	4.1	100
Auvergne	83	11.8	10.1	5.0	17.8	3.7	1.0	5.9	7.9	2.3	4.0	9.8	7.1	9.4	4.4	100
Languedoc-Roussillon	91	12.1	10.9	5.7	19.9	3.2	1.0	6.1	7.3	2.6	4.3	10.4	4.6	8.1	3.9	100
Provence-Alpes-Côte d'Azur	93	11.7	10.4	5.9	19.8	3.2	1.0	5.7	6.9	2.6	3.8	10.1	5.4	10.3	3.4	100
Corse	94	12.6	11.9	6.7	19.4	3.4	1.1	7.3	7.4	2.7	4.1	8.2	4.5	8.1	2.7	100
France métropolitaine (1)		11.9	10.3	5.1	18.7	3.2	1.0	5.9	7.2	2.4	3.6	10.6	6.4	10.2	3.6	100
France (2)		14.3	21.6	5.2	12.1	1.8	5.8	9.8	6.8	3.4	2.2	5.4	4.1	5.7	1.7	100

Note : répartition (%) territoriale du chiffre d'affaires, selon le type de produit, par regroupement de classes de la nomenclature COICOP. 01.1.1 : Pain et céréales ; 01.1.2 : Viande ; 01.1.3 : Poissons et crustacés ; 01.1.4 : Lait, fromage et œufs ; 01.1.5 : Huiles et graisses ; 01.1.6 : Fruits ; 01.1.7 Légumes ; 01.1.8 : Sucre, confitures, chocolat, confiserie et produits glacés ; 01.1.9 : Sel, épices, sauces et produits alimentaires non ailleurs ; 01.2.1 : Café, thé et cacao ; 01.2.2 : Autres boissons non alcoolisées ; 02.1.1 : Alcools ; 02.1.2 : Vins, cidres et champagne ; 02.1.3 : Bières. Calcul des auteurs sur la base des données de caisse de la semaine de référence (avril 2013), y compris pour (1). (2) répartition France entière, Comptabilité nationale (tableaux détaillés de consommation des ménages pour 2013).
Source : Insee, données de caisse 2013.

retenu permet d'estimer les écarts moyens de prix entre zones géographiques. Formellement, on postule que le prix p_i répond à un processus générateur de la forme :

$$\log(p_i) = c + \sum_{\ell=1}^{L} \alpha_\ell \cdot \mathbf{1}_{(ean_i=\ell)} + \sum_{z=1}^{Z} \beta_z \cdot \mathbf{1}_{(zone_i=z)} + \varepsilon_i \quad (1)$$

où **1** désigne une variable indicatrice valant 1 si la condition figurant entre parenthèses en indice est vraie et 0 sinon, ean_i est le numéro de code-barres de l'observation i et $zone_i$ est la zone géographique à laquelle appartient l'observation i. ε_i est un aléa centré. Dans ce modèle, les coefficients c, α_ℓ ($\ell \in \{1, \ldots, L\}$, L est le nombre de codes-barres pris en compte) et β_z ($z \in \{1, \ldots, Z\}$, Z est le nombre de zones géographiques prises en compte) sont inconnus. On les estime par moindres carrés. Les ratios[11] des coefficients α_ℓ s'interprètent comme les rapports de prix moyens associés aux codes-barres considérés. Les rapports des coefficients β_z traduisent les rapports de prix moyens entre zones géographiques, à produits donnés (repérés par leurs codes-barres). Ces coefficients, estimés par moindres carrés, correspondent à des indices de prix hédoniques (Triplett, 2006 ; Diewert, 2003 ; Silver & Heravi, 2005).

La forme des estimateurs obtenus est détaillée dans l'encadré. On constate que l'estimateur obtenu fait naturellement intervenir les différences de structure de consommation entre régions, par l'intermédiaire des pondérations utilisées. La pondération la plus naturelle, de ce point de vue, est par le chiffre d'affaires du produit dans le point de vente considéré. Le modèle de référence fait donc intervenir une pondération par le chiffre d'affaires. La pondération unitaire fait *de facto* intervenir une structure, assez proche de celle par le chiffre d'affaires, puisqu'elle se fonde sur le nombre de transactions pour le produit et le point de vente considérés. L'approche alternative par la pondération unitaire est donc utilisée afin d'examiner la robustesse des résultats vis-à-vis de la pondération de référence.

11. Le rapport des exponentielles de ces coefficients pour être précis (voir infra*).*

ENCADRÉ – **Structure des estimateurs hédoniques**

L'estimateur des moindres carrés (1) peut ou non être pondéré. Concrètement, il y a deux options possibles : soit on utilise des pondérations s'apparentant aux chiffres d'affaires ω_i, soit on ne pondère pas les observations élémentaires. Afin de bien évaluer les conséquences du choix que nous faisons en termes de pondérations, il est utile d'examiner la forme des estimateurs que nous obtenons pour les coefficients β_z. Pour cela, nous supposons pour plus de simplicité que l'estimation est réalisée en deux étapes[(a)] : une première étape dans laquelle les coefficients α_ℓ sont estimés. Puis, dans une seconde étape, les coefficients β_z sont estimés (conditionnellement aux estimateurs $\hat{\alpha}_\ell$ des α_ℓ obtenus en première étape). Évidemment, en procédant de la sorte, nous n'obtenons pas l'estimateur des moindres carrés que nous obtiendrions si les vecteurs ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) étaient estimés simultanément, mais les limites en probabilité des deux estimateurs en deux étapes sont les mêmes que celles de l'estimateur en une étape[(b)]. L'intérêt de procéder en deux étapes est que l'on peut aisément dériver la forme de $\hat{\beta}$. En effet, soit $\tilde{p}_i$ la variable p_i corrigée de la première étape :

$$\log(\tilde{p}_i) = \log(p_i) - c - \sum_{\ell=1}^{L} \hat{\alpha}_\ell \cdot \mathbf{1}_{(ean_i=\ell)} \quad (2)$$

alors la deuxième étape consiste à régresser $\log(\tilde{p}_i)$ sur les vecteurs ligne $\boldsymbol{x}_i$ comprenant Z colonnes, dont $Z-1$ sont nulles, et la seule non nulle est égale à 1 :

$$\log(\tilde{p}_i) = \boldsymbol{x}_i \cdot \boldsymbol{\beta} + \nu_i \quad (3)$$

L'estimateur des moindres carrés $\hat{\boldsymbol{\beta}}$ est classiquement solution de l'équation (ici en version pondérée ; pour une version non pondérée, il suffit de poser $\omega_i = 1$) :

$$\left(\sum_{i\in I} \omega_i \boldsymbol{x}_i{}' \cdot \boldsymbol{x}_i\right) \cdot \hat{\boldsymbol{\beta}} = \sum_{i\in I} \omega_i \boldsymbol{x}_i{}' \cdot \log(\tilde{p}_i)$$

soit, en regroupant par modalité de zone[(c)] :

$$Diag\begin{pmatrix} \sum_{i\in z_1} \omega_i \\ \vdots \\ \sum_{i\in z_z} \omega_i \end{pmatrix} \cdot \hat{\boldsymbol{\beta}} = \begin{pmatrix} \sum_{i\in z_1} \omega_i \log(\tilde{p}_i) \\ \vdots \\ \sum_{i\in z_z} \omega_i \log(\tilde{p}_i) \end{pmatrix}$$

et finalement, pour la zone k considérée (aussi notée z_k) :

$$\exp(\hat{\beta}_k) = \left\{\prod_{i\in z_k} \tilde{p}_i^{\omega_i}\right\}^{1/\sum_{i\in z_k}\omega_i} \quad (4)$$

Il en découle que, pour les zones k et j, nous avons :

$$exp(\hat{\beta}_j - \hat{\beta}_k) = \frac{\left\{\prod_{i\in z_j} \tilde{p}_i^{\omega_i}\right\}^{1/\sum_{i\in z_j}\omega_i}}{\left\{\prod_{i\in z_k} \tilde{p}_i^{\omega_i}\right\}^{1/\sum_{i\in z_k}\omega_i}} \quad (5)$$

On note que ce ratio correspond à un rapport de prix moyens[(d)] (i.e. ratio de valeurs unitaires). On observe que cet indice d'écart de niveau de prix entre zones prend en compte les structures locales de consommation puisque, tant au numérateur qu'au dénominateur, chaque produit pèse dans l'indice en proportion de son poids dans la dépense de consommation locale.

(a) Cette décomposition en deux étapes est uniquement proposée ici dans le but d'expliciter la forme de l'indice obtenu. En pratique, nous faisons un calcul en une seule étape fondé sur le modèle (1).
(b) Sous les mêmes hypothèses de convergence, notamment d'orthogonalité de l'aléa et des variables explicatives.
(c) Diag désigne la matrice diagonale dont la diagonale coïncide avec le vecteur en argument.
(d) En moyenne géométrique, à interpréter comme étant calculé à code-barres fixé, identique pour le numérateur et le dénominateur, en raison du conditionnement par l'EAN dans les étapes (1) et (2).

À ce stade, il convient de préciser les conditions sous lesquelles l'estimateur $\hat{\beta}_k$ est non biaisé. En tant qu'estimateur du coefficient β_k de l'équation (1), le coefficient est non biaisé dès que les conditions d'orthogonalité des variables explicatives et de l'aléa ε_i (ou v_i dans le cas de la régression de deuxième étape) sont assurées. On fait l'hypothèse ici que c'est bien le cas. En revanche, la statistique $\exp(\hat{\beta}_k)$ n'est pas un estimateur sans biais de $\exp(\beta_k)$. En effet, à partir de l'expression de l'estimateur des moindres carrés (équation 1 ou 3), on montre[12] que :

$$E\left[\frac{p_i}{p_\ell}\middle| i \in z_k, \ell \in z_j, ean_i = ean_j\right] = \exp\left(\beta_k - \beta_j\right)\left[1+\sigma^2\right] \quad (6)$$

où σ^2 est la variance des ε_i qu'on supposera désormais de variance identique. Nous utiliserons donc cette correction pour calculer les rapports de prix.

Résultats

Les écarts observés en avril 2013

Dans cette partie, nous présentons les résultats fondés sur des régressions du type du modèle (1) pour une semaine de données en avril 2013 (troisième semaine du mois). D'un point de vue pratique, pour toutes les régressions réalisées, on ne conserve que 5 000 références de codes-barres par enseigne. On choisit, parmi les références vendues de l'enseigne, les 5 000 principales en termes de chiffre d'affaires. En effet, le modèle hédonique (1) est fondé sur des indicatrices de codes-barres. Celles-ci ne sont pas explicitement estimées (elles sont réduites algébriquement dans l'équation normale), mais un trop grand nombre de références conduit à une équation normale trop lourde à traiter. Différents tests ont été menés pour examiner les conséquences de cette restriction. L'expérience montre que retenir 3 000 ou 5 000 références par enseigne ne conduit pas, sur les indicatrices géographiques, à des résultats notablement différents. Au final, la réunion, pour l'ensemble des enseignes de la base, des 5 000 principaux codes-barres les concernant, conduit à considérer 13 098 codes-barres dans les régressions. Ce nombre étant nettement plus élevé que les 5 000 références conservées par enseigne, une fraction significative des codes-barres est donc spécifique[13] aux enseignes (marques de distributeur). Compte-tenu de cette restriction, la base de calcul comprend 7.3 millions d'enregistrements correspondant aux croisements [point de vente × codes-barres retenus]. En termes de chiffre d'affaires, la restriction effectuée conduit à conserver 74 % de l'information contenue dans la base d'origine présentée dans la section sur les données.

Le tableau 5 indique, par type de produits et pour la base de données restreinte aux 5 000 principaux

12. Par exemple par une Δ-méthode ou bien en faisant des hypothèses sur la distribution normale des aléas dans l'équation (1). E *désigne l'espérance mathématique (notation conditionnelle).*

13. Si chaque code-barres était vendu dans toutes les enseignes, la réunion des 5 000 principaux codes-barres d'enseignes comprendrait précisément 5 000 codes-barres.

Tableau 5
Répartition des familles IRI et des codes-barres par regroupement de la nomenclature COICOP

Code COICOP	Libellé COICOP	Nombre de familles	Nombre de codes-barres
0111	Pain	47	2 200
0112	Viande	19	1 479
0113	Poissons et crustacés	22	848
0114	Lait, fromage et œufs	23	1 830
0115	Huiles et graisses	6	300
0116	Fruits	15	252
0117	Légumes	31	1 117
0118	Sucre, confitures, chocolat, confiseries et produits glacés	29	1 098
0119	Sel, épices sauces et autres produits alimentaire	35	564
0121	Café, thé et cacao	10	409
0122	Autres boissons non alcoolisées	17	876
0211	Alcools	12	361
0212	Vins, cidres et champagne	21	1 535
0213	Bières	1	229
Total		288	13 098

Lecture : la base de données comprend 47 familles de produits IRI appartenant au regroupement COICOP 0111 (Pains). 2 200 références codes-barres s'y rapportent dans la base de données examinée.
Source : Insee, données de caisse 2013.

codes-barres par enseigne, le nombre de famille de produits IRI[14] qui s'y rattachent, ainsi que le nombre de références code-barres correspondant. Grossièrement, une famille IRI correspond à un type de produit dont le grain est approximativement de même finesse que celui des variétés de produits suivies dans l'IPC (Insee, 1998). Pour mémoire, 327 variétés sont suivies dans l'IPC métropolitain au titre de l'alimentaire industriel en 2013. Ce chiffre est effectivement comparable au nombre de familles IRI qui, sur le même champ, est de 288. Dans la base de données étudiée, le nombre de références code-barres correspondant est, comme vu plus haut, de 13 098.

Le tableau 6 donne les résultats d'estimation des indices d'écart de niveaux de prix dans l'alimentaire industriel pour les régions administratives métropolitaines, calculés à l'aide de la base de données de caisse. On constate, tout d'abord, que les écarts sont relativement peu dispersés : 5.5 à 8 points de pourcentage selon que l'on pondère ou non les observations par leurs chiffres d'affaires. La dispersion est plus importante lorsqu'on considère les indices non pondérés plutôt que les indices pondérés. Cela suggère que les produits pesant davantage dans le budget des consommateurs connaissent une dispersion spatiale de prix plus faible que les autres produits. Il est aussi remarquable de constater que l'ordre des régions classées par niveau d'écart moyen de prix n'est pas modifié selon que l'on pondère ou non les observations par le chiffre d'affaires.

Sur un plan géographique, les résultats dessinent des dominantes régionales : un grand Centre-Ouest de la France dans lequel les niveaux de prix sont environ 3 % plus bas qu'en Île-de-France ; puis une partie comprenant les régions plus rurales du Centre, celles du Nord de la France ainsi que l'Aquitaine dans lesquelles les prix de l'alimentaire industriel sont en moyenne de 2 % inférieurs à ceux de l'Île-de-France ; les régions plus industrielles et urbaines de l'Est et du Sud présentent des niveaux de prix alimentaires 1 % plus bas

14. Compagnie privée qui développe un catalogue, utilisé par l'Insee dans le cadre de l'expérience pilote, de caractéristiques de produits référencés par codes-barres.

Tableau 6
Indices d'écart de niveau de prix entre la région parisienne et les autres régions

Région	Code	Estimation		
		Pondérée	Non pondérée	Pondérée avec E.F. d'enseigne
Bretagne	53	**96.7**	95.4	97.1
Pays de la Loire	52	**97.0**	96.1	97.6
Centre	24	**97.6**	96.8	97.9
Limousin	74	**97.8**	96.5	98.0
Poitou-Charentes	54	**97.4**	96.6	98.2
Basse-Normandie	25	**97.9**	96.8	98.2
Auvergne	83	**98.2**	97.2	98.4
Haute-Normandie	23	**98.1**	97.5	98.4
Midi-Pyrénées	73	**98.3**	97.2	98.4
Nord-Pas-de-Calais	31	**97.9**	97.1	98.6
Bourgogne	26	**97.7**	96.9	98.6
Picardie	22	**98.2**	97.4	98.6
Aquitaine	72	**98.2**	97.3	98.6
Franche-Comté	43	**97.9**	97.1	98.7
Champagne-Ardenne	21	**98.1**	97.4	98.7
Alsace	42	**98.9**	98.5	98.9
Lorraine	41	**98.6**	98.0	99.0
Languedoc-Roussillon	91	**98.6**	98.0	99.2
Rhône-Alpes	82	**98.9**	98.2	99.3
Provence-Alpes-Côte d'Azur	93	**99.2**	98.9	99.9
Île-de-France	11	**100** (Réf.)	100 (Réf.)	100 (Réf.)
Corse	94	**102.1**	103.5	102.8

Lecture : au sens de l'estimation dans laquelle les observations sont pondérées par leur chiffre d'affaires, les prix sont en moyenne 3.3 % plus bas en Bretagne qu'en Île-de-France. Au sens de l'estimation dans laquelle les observations sont pondérées unitairement, les prix sont en moyenne 4.4 % plus bas en Bretagne qu'en Île-de-France. Les indicatrices de zones résultent d'une régression de type (1) dans laquelle les zones sont les anciennes régions administratives. La dernière colonne se réfère à un calcul équivalent à celui réalisé pour la première colonne (i. e. pondéré), dans lequel un effet-fixe enseigne a été ajouté. Les résultats obtenus sont corrigés conformément à la formule (6) et transformés en indices par une multiplication par 100. La variance estimée de l'aléa est de 0.004. Calcul effectué sur 7.3 millions d'enregistrements. L'écart-type moyen sur les indices présentés est de 0.02 point d'indice.
Source : Insee, données de caisse 2013.

qu'en Île-de-France. Enfin, les prix en Corse sont 2 % plus élevés qu'en Île-de-France.

Afin de comparer aux résultats « historiques » présentés au tableau 1, le tableau 7 regroupe les indices d'écarts de prix alimentaires industriels entre les grandes agglomérations de métropole et l'agglomération parisienne. Pour comparer ces résultats à ceux du tableau 1, il convient de rappeler que les champs économiques et géographiques ainsi que les méthodes de calcul ne sont pas rigoureusement homogènes. Une partie des écarts constatés entre agglomérations et leurs évolutions au cours du temps intègrent vraisemblablement des biais liés à l'inhomogénéité de champ et de méthode. Néanmoins, il est quand même intéressant d'examiner les résultats obtenus.

Pour les agglomérations comme pour les régions, on constate (cf. tableaux 6 et 7) que les écarts de niveaux de prix estimés par régression non pondérée sont un peu plus importants que ceux calculés par régression pondérée. Hors Corse[15], les écarts de prix se situent dans une fourchette de 3.7 à 4.4 points de pourcentage selon qu'on pondère ou non les observations. Par rapport à l'agglomération parisienne dans laquelle les prix sont les plus élevés, les agglomérations les moins chères (parmi les grandes agglomérations) pour l'alimentaire industriel sont Nantes, Rennes, Orléans, Rouen et Lille. Remarquablement, c'était déjà le cas en 1989 (Insee, Division prix de détail 1990) et en 1985 (Mineau, 1987) – cf. tableau 1. La distance avec le schéma de 1977 (Baraille, 1978) est légèrement plus marquée.

Par rapport aux écarts mis en évidence entre régions, ceux constatés entre grandes agglomérations sont un peu plus accentués. Par exemple, en référence à une zone quasi-comparable

15. Non présentée dans le tableau des agglomérations (tableau 7) en raison d'un nombre de points de vente trop restreint dans la base de données de caisse.

Tableau 7
Écarts de niveau de prix entre l'agglomération de Paris et les principales autres agglomérations métropolitaines

Agglomération	Estimation	
	Pondérée	Non pondérée
Agglomération parisienne	100 (Réf.)	100 (Réf.)
Lyon	98.6	97.7
Marseille	98.9	98.4
Bordeaux	97.9	97.0
Rennes	96.5	95.6
Reims	97.9	97.6
Rouen	97.1	96.6
Strasbourg	99.1	98.7
Lille	97.3	96.5
Orléans	97.1	95.6
Limoges	nd	nd
Ajaccio-Bastia	nd	nd
Clermont-Ferrand	98.4	97.3
Toulouse	98.0	96.7
Dijon	nd	nd
Nantes	96.3	95.9
Nancy	98.4	97.7
Poitiers	nd	nd
Montpellier	97.9	97.1
Limoges	96.6	95.8
Ajaccio-Bastia	101.5	102.3
Dijon	97.1	96.5
Poitiers	97.7	97.0

Lecture : au sens de l'estimation dans laquelle les observations sont pondérées par leur chiffre d'affaires, les prix sont en moyenne 1.4 % plus bas à Lyon qu'à Paris. Au sens de l'estimation dans laquelle les observations sont pondérées unitairement, les prix sont en moyenne 2.3 % plus bas à Lyon qu'à Paris. Les indicatrices de zones résultent d'une régression de type (1) dans laquelle les zones sont les agglomérations (unités urbaines). Les résultats obtenus sont corrigés conformément à la formule (6) et transformés en indices par une multiplication par 100. La variance estimée de l'aléa est de 0.004. Calcul effectué sur 7.3 millions d'enregistrements. L'écart-type moyen sur les indices présentés est de 0.10 point d'indice.
Source : Insee, données de caisse 2013.

(l'agglomération parisienne ou l'Île-de-France selon le cas), l'indice (pondéré) de Montpellier est de 97.9 tandis que celui du Languedoc-Roussillon est de 98.6. De même, celui de Lille est de 97.3 tandis que celui du Nord-Pas-de-Calais est de 97.9. Cette situation[16] pourrait être liée au fait que la concurrence est probablement plus forte au sein des marchés locaux de grandes agglomérations, ce qui aurait tendance à tirer les prix vers le bas.

Cette règle souffre toutefois de deux exceptions parmi les grandes agglomérations : il s'agit de Strasbourg dont l'indice est de 99.1 tandis que celui de l'Alsace est de 98.9 et Clermont-Ferrand dont l'indice est de 98.4 tandis que celui de l'Auvergne est de 98.2. Dans ces deux cas, les écarts ne sont néanmoins pas significatifs.

Comme évoqué en introduction et plus haut, la représentativité de l'échantillon de données par rapport à la distribution spatiale des prix peut être altérée du fait du nombre limité d'enseignes ayant fourni leurs données à l'Insee en 2013. Ainsi, il est possible que la sélection des enseignes de l'échantillon soit corrélée à la dimension régionale sur laquelle sont estimées les statistiques proposées. C'est le cas, par exemple, si une enseigne de l'échantillon dont la politique de prix diffère des autres (disons par exemple que ses prix sont systématiquement moins élevés) est, du fait de la sélection, surreprésentée dans une région et pas ailleurs. Dans ce cas, l'estimation du niveau des prix dans la région concernée par la surreprésentation est biaisée (à la baisse pour l'exemple) par rapport aux autres régions.

Pour démontrer l'absence ou l'existence d'un tel biais et pour l'évaluer, il faudrait disposer d'un jeu complet de données relatif à l'ensemble des enseignes. Avec l'échantillon limité dont on dispose, s'il n'est pas possible de procéder à une étude définitivement concluante sur ce point, il est toutefois possible d'examiner la cohérence de certains résultats avec l'hypothèse de représentativité du sous-échantillon étudié. Le premier résultat utile à cet égard a déjà été présenté dans le tableau 3 qui montre que la distribution régionale des points de vente respecte la distribution de la population et donc, vraisemblablement, la dépense de consommation des ménages dans l'alimentaire. Un autre résultat utile montre l'intérêt d'ajouter dans l'équation (1) des indicatrices d'enseignes. Si, par exemple, un indice régional est significativement modifié dans ce second calcul par rapport au calcul de référence, c'est le signe que le niveau régional des prix tient pour partie aux enseignes représentées dans le réseau local des supermarchés, dans le sous-échantillon étudié. Il est possible, dans ces conditions, que la portée des résultats obtenus soit, au fond, limitée au seul échantillon considéré. Nous avons procédé à un tel calcul dont les résultats, en termes d'indices régionaux pondérés par les ventes, sont présentés au tableau 6 en dernière colonne. Ces résultats sont à comparer avec ceux du calcul de référence (en gras dans ce même tableau). Il s'avère que les indices régionaux peuvent être assez nettement modifiés, jusqu'à 0.8 points pour la Bourgogne et la Franche-Comté. Ceci étant, les constats principaux, en particulier la hiérarchie des prix entre la Corse, l'Île-de-France et les autres régions métropolitaines, de même que l'ordre de grandeur des écarts, demeurent.

Finalement, si des « effets enseignes » existent, leur impact dans les indices locaux étant perceptible, la généralisation, à l'ensemble de la consommation alimentaire dans la grande distribution, des principaux enseignements tirés du sous-échantillon étudié est raisonnablement accréditée par les différents tests de robustesse réalisés.

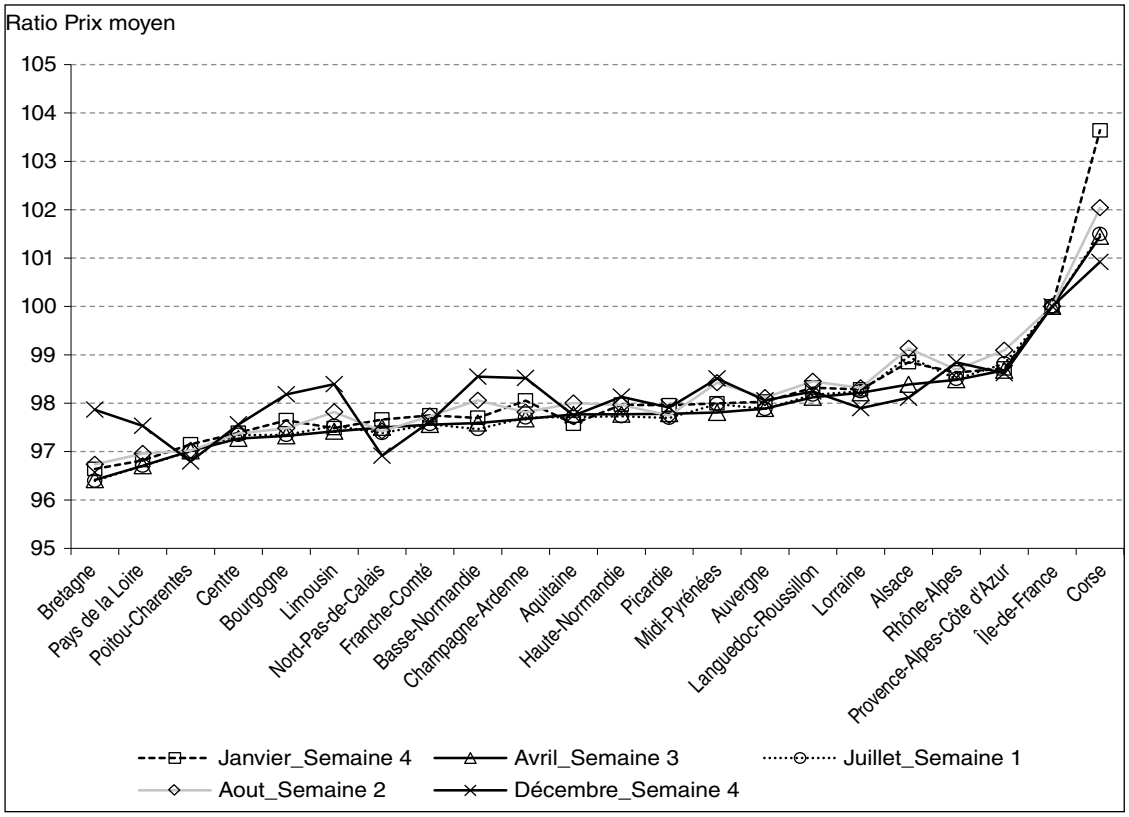
Sensibilité des résultats au choix de la semaine de calcul

Afin de tester la robustesse des résultats obtenus, nous examinons maintenant le comportement des écarts régionaux de niveaux de prix lorsque la semaine retenue change. Pour cela, nous généralisons l'analyse précédente à quatre autres semaines de l'année 2013 relativement typiques quant à leur situation par rapport aux soldes et aux fêtes à fort impact sur les achats de consommation : quatrième semaine de janvier (peu après les fêtes de fin d'année et en pleine période de soldes d'hiver), première semaine de juillet (début des vacances d'été), deuxième semaine d'août (fin des vacances d'été) et quatrième semaine de décembre (fêtes de fin d'année). Ces semaines sont comparées à la troisième semaine d'avril, précédemment étudiée et servant de référence à la comparaison.

La figure suivante montre les écarts de niveaux de prix entre l'Île-de-France et les autres régions pour les 5 semaines étudiées, les régions étant ordonnées en abscisse selon leur rang en termes de niveau de prix observé lors de la semaine de référence d'avril. On constate que les écarts sont

16. Pour l'interprétation, on fait ici l'hypothèse, raisonnable compte tenu de son poids, que le niveau de prix de l'agglomération parisienne est aussi celui de l'Île-de-France. Par conséquent, les écarts d'indices des villes de province et de leur région sont liées à des différences locales entre les villes et leur région d'appartenance et non à d'éventuelles différences de prix entre l'unité urbaine de Paris et sa région.

Figure
Écart de niveaux de prix régionaux par rapport à l'Île-de-France



Note : référence 100 pour l'Île-de-France pour chacune des semaines d'étude. L'ordre des régions figurant en abscisses est croissant selon le niveau d'indice constaté en avril 2013.
Source : Insee, données de caisse 2013.

extrêmement voisins d'une semaine à l'autre. Il existe toutefois deux exceptions à cette proximité. D'une part, le niveau des prix en Corse est relativement plus élevé en janvier que lors des autres semaines examinées. D'autre part, on note une structure de prix régionale assez nettement modifiée lors de la dernière semaine de décembre ce qui est l'effet probable de la spécificité des produits vendus à cette époque et des mouvements de population importants lors des fêtes qui modifient la structure géographique des marchés.

Au final, cette analyse de robustesse tend à confirmer le caractère essentiellement structurel des écarts géographiques de niveaux de prix. Elle montre également la richesse des lots de données de caisse qui permettent d'estimer avec une grande précision des indices de prix sur des pas géographiques ou temporels inaccessibles aux moyens d'enquête classiques.

* *
*

Cette étude présente un exemple d'exploitation des données de caisse pour établir des mesures de différences de niveaux de prix entre territoires métropolitains dans l'alimentaire et les boissons alcoolisées ou non. Naturellement, compte-tenu du champ couvert par les données de caisse utilisées, la portée de ces résultats reste limitée et leur généralisation à toute la consommation alimentaire des ménages métropolitains est sujette à discussion. D'abord parce que les enseignes concourant à l'expérience pilote menée par l'Insee en 2013 sont relativement peu nombreuses (même si elles contribuent à hauteur de 30 % du chiffre d'affaires de la grande distribution), ensuite parce que la répartition de leurs points de vente sur l'ensemble du territoire métropolitain n'est vraisemblablement pas parfaitement représentative de la géographie des lieux de consommation des ménages. Au niveau régional néanmoins, les résultats présentés au tableau 3 suggèrent que l'échantillon étudié ne souffre pas de déséquilibre spatial évident par rapport à la distribution de la population.

Par rapport aux travaux existants qui ont été présentés dans la première section, le fait de

mesurer des écarts de niveau de prix conditionnellement à un identifiant univoque du produit – le code-barres en l'occurrence – affermit certainement les constats. De même, l'ensemble des produits pris en compte dans le calcul des différences de niveaux améliore la précision, en raison de leur nombre considérable, et permet de couvrir la quasi-exhaustivité du champ de l'alimentaire et des boissons alcoolisées ou non, référencé par code-barres, tandis que les études antérieures devaient se contenter de s'appuyer sur des représentants de produits dont la représentativité n'était pas évidente à justifier. Au final, ces travaux apportent d'ores et déjà une information importante et très crédible sur les écarts de niveaux de prix dans l'alimentaire, notamment s'agissant d'agglomérations de taille importante. Ils établissent que la dispersion est relativement faible, comme les travaux historiques l'avaient montré, et qu'elle n'a probablement que peu évolué sur près de 40 ans. □

BIBLIOGRAPHIE

Anxionnaz, I. & Mothe, A. (2000). Les comparaisons spatiales de prix au sein du territoire français : historique et développements à prévoir. *Courrier des statistiques*, 98-96, 11–16.
https://www.insee.fr/fr/metadonnees/source/fichier/ipc_courrierstat_95_comparaisons_spatiales.pdf

Baraille, J. (1978). Les prix dans les grandes villes de France. *Economie et Statistique*, 106, 17–20.
https://www.persee.fr/doc/estat_0336-1454_1978_num_106_1_3004

Baraille, J. P. & Bobin, M. F. (1981). Les écarts de prix à l'intérieur de la métropole. *Économie et Statistique*, 130, 61–66.
https://www.persee.fr/doc/estat_0336-1454_1981_num_130_1_4453

Berthier, J., Lhéritier, J. & Petit, G. (2010). Comparaison des prix entre la métropole et les DOM en 2010. *Insee Première* N° 1304.
https://www.insee.fr/fr/statistiques/1287446

Clé, E., Sauvadet, L., Jaluzot, L., Malaval, F. & Rateau, G. (2016). En 2015, les prix en région parisienne dépassent de 9 % ceux de la province. *Insee Première* N° 1590.
https://www.insee.fr/fr/statistiques/1908158

de Haan, J. & van der Grient, H. (2011). Eliminating Chain Drift in Price Indexes Based on Scanner Data. *Journal of Econometrics*, 161(1), 36–46.
https://ideas.repec.org/a/eee/econom/v161y2011i1p36-46.html

Deaton, A. & Heston, A. (2010). Understanding PPPs and PPP-based national accounts. *American Economic Journal: Macroeconomics*, 2(4), 1–35.
https://doi.org/10.1257/mac.2.4.1

Diewert, E. (2003). Hedonic Regressions: A Consummer Theory Approach. *In*: Feenstra, R. C. & Shapiro, M. D. (Eds), *Scanner Data and Price Indexes*. Chicago: University of Chicago Press.

Insee, Division prix de détail (1990). Les prix dans 23 agglomérations en 1989. *Insee Première* N° 69.
https://www.epsilon.insee.fr/jspui/bitstream/1/10075/1/ip69.pdf

Fesseau, M., Passeron, V. & Vérone, M. (2008). Les prix sont plus élevés en Île-de-France qu'en province. *Insee Première* N° 1210.
https://www.insee.fr/fr/statistiques/1281287

Guglielmetti, F. (1996). Les prix en Corse : entre Marseille et Paris. *Insee Première* N° 442.
https://www.epsilon.insee.fr/jspui/bitstream/1/890/1/ip442.pdf

Insee (1998). Pour comprendre l'indice des prix. *Insee-méthodes* N° 81-82.
https://www.insee.fr/fr/metadonnees/source/fichier/Indice_des_prix.pdf

Insee, Division prix de détail (1990). Les prix dans 23 agglomérations en 1989. *Insee Première* N° 69.
http://www.epsilon.insee.fr/jspui/bitstream/1/890/1/ip442.pdf

Mineau, B. (1987). Les comparaisons de prix entre agglomérations françaises. *Courrier des statistiques*, 44, 21–24.

Nicolaï, M. P. (2010). Enquête de comparaison spatiale des prix Corse-Continent 2010. *Quant'île – Insee Corse* N° 12.
https://www.insee.fr/fr/statistiques/fichier/1378434/quantile12.pdf

Piccard, H. (1972). Situation relative des prix de détail dans les agglomérations de plus de 20 000 habitants en octobre 1971. *Économie et Statistique*, 37, 35–38.
https://doi.org/10.3406/estat.1972.1242

Reinsdorf, M. (1999). Using Scanner Data to Construct CPI Basic Component Indexes. *Journal of Business & Economic Statistics*, 17, 152–160.
https://www.jstor.org/stable/1392470

Silver, M. & Heravi, S. (2005). A failure in the measurement of inflation: Results from a hedonic and matched experiment using scanner data. *Journal of Business & Economic Statistics*, 23(3), 269–281.
https://doi.org/10.1198/073500104000000343

Triplett, J. (2006). Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. *Documents de travail de l'OCDE sur la science, la technologie et l'industrie* N° 2004/09.
http://dx.doi.org/10.1787/643587187107

N° 507-508 – 2019

MÉLANGE / *VARIA*

N° 505-506 – 2018

BIG DATA ET STATISTIQUES 1ère PARTIE / *BIG DATA AND STATISTICS PART 1*

PRÉVISION « IMMÉDIATE » / *NOWCASTING*

DONNÉES DE TÉLÉPHONIE MOBILE / *MOBILE PHONES DATA*

DONNÉES ET MÉTHODES / *DATA AND METHODS*

BIG DATA ET STATISTIQUE PUBLIQUE / *BIG DATA AND OFFICIAL STATISTICS*

Economie et Statistique / Economics and Statistics

Objectifs généraux de la revue

Economie et Statistique / Economics and Statistics publie des articles traitant de tous les phénomènes économiques et sociaux, au niveau micro ou macro, s'appuyant sur les données de la statistique publique ou d'autres sources. Une attention particulière est portée à la qualité de la démarche statistique et à la rigueur des concepts mobilisés dans l'analyse. Pour répondre aux objectifs de la revue, les principaux messages des articles et leurs limites éventuelles doivent être formulés dans des termes accessibles à un public qui n'est pas nécessairement spécialiste du sujet de l'article.

Soumissions

Les propositions d'articles, en français ou en anglais, doivent être adressées à la rédaction de la revue (redaction-ecostat@insee.fr), en format MS-Word. Il doit s'agir de travaux originaux, qui ne sont pas soumis en parallèle à une autre revue. Un article standard fait environ 11 000 mots (y compris encadrés, tableaux, figures, annexes et bibliographie, non compris éventuels compléments en ligne). Aucune proposition initiale de plus de 12 500 mots ne sera examinée.

La soumission doit comporter deux fichiers distincts :

- Un fichier d'une page indiquant : le titre de l'article ; le prénom et nom, les affiliations (maximum deux), l'adresse e-mail et postale de chaque auteur ; un résumé de 160 mots maximum (soit environ 1 050 signes espaces compris) qui doit présenter très brièvement la problématique, indiquer la source et donner les principaux axes et conclusions de la recherche ; les codes JEL et quelques mots-clés ; d'éventuels remerciements.
- Un fichier anonymisé du manuscrit complet (texte, illustrations, bibliographie, éventuelles annexes) indiquant en première page uniquement le titre, le résumé, les codes JEL et les mots-clés.

Les propositions retenues sont évaluées par deux à trois rapporteurs (procédure en « double-aveugle »). Les articles acceptés pour publication devront être mis en forme suivant les consignes aux auteurs (accessibles sur https://www.insee.fr/fr/information/2410168). Ils pourront faire l'objet d'un travail éditorial visant à améliorer leur lisibilité et leur présentation formelle.

Publication

Les articles sont publiés en français dans l'édition papier et simultanément en français et en anglais dans l'édition électronique. Celle-ci est disponible, en accès libre, sur le site de l'Insee, le jour même de la publication ; cette mise en ligne immédiate et gratuite donne aux articles une grande visibilité. La revue est par ailleurs accessible sur le portail francophone Persée, et référencée sur le site international Repec et dans la base EconLit.

Main objectives of the journal

Economie et Statistique / Economics and Statistics publishes articles covering any micro- or macro- economic or sociological topic, either using data from public statistics or other sources. Particular attention is paid to rigor in the statistical approach and clarity in the concepts and analyses. In order to meet the journal aims, the main conclusions of the articles, as well as possible limitations, should be written to be accessible to an audience not necessarily specialist of the topic.

Submissions

Manuscripts can be submitted either in French or in English; they should be sent to the editorial team (redaction-ecostat@insee.fr), in MS-Word format. The manuscript must be original work and not submitted at the same time to any other journal. The standard length of an article is of about 11,000 words (including boxes if needed, tables and figures, appendices, list of references, but not counting online complements if any). Manuscripts of more than 12,500 words will not be considered.

Submissions must include two separate files:

- A one-page file providing: the title of the article; the first name, name, affiliation-s (at most two), e-mail et postal addresses of each author; an abstract of maximum 160 words (about 1050 characters including spaces), briefly presenting the question(s), data and methodology, and the main conclusions; JEL codes and a few keywords; acknowledgements.
- An anonymised manuscript (including the main text, illustrations, bibliography and appendices if any), mentioning only the title, abstract, JEL codes and keywords on the front page.

Proposals that meet the journal objectives are reviewed by two to three referees ("double-blind" review). The articles accepted for publication will have to be presented according to the guidelines for authors (available at https://www.insee.fr/en/information/2591257). They may be subject to editorial work aimed at improving their readability and formal presentation.

Publication

The articles are published in French in the printed edition, and simultaneously in French and in English in the online edition. The online issue is available, in open access, on the Insee website the day of its publication; this immediate and free online availability gives the articles a high visibility. The journal is also available online on the French portal Persée, and indexed in Repec and EconLit.

N° 509 - 2019

Economie ET Statistique

Economics AND Statistics

Au sommaire du prochain numéro :
Numéro spécial 50ème anniversaire

Forthcoming:
Special Issue 50th Anniversary

ISBN 978-2-11-151240-5 - ISSN 0336-1454 - ECO 509
Parution septembre 2019 - PRIX : 9,10 €