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

* *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. □