STATBEL
La Belgique en chiffres

L'utilisation des scanner data des supermarchés dans l'indice des prix à la consommation

- Ken Van Loon et Dorien Roels -

n°01

ANALYSE

12.2017

economie
SPF Economie, P.M.E., Classes moyennes et Energie

.be

L'utilisation des scanner data des supermarchés dans l'indice des prix à la consommation

Dorien Roels, Ken Van Loon[1]

[1] Statisticiens à Statbel (Direction générale Statistique - Statistics Belgium)

ABSTRACT

La présente analyse explique l'utilisation des scanner data des supermarchés dans l'indice des prix à la consommation. En quoi consistent les scanner data? À quoi ressemblent ces données? Et comment la DG Statistique – Statistics Belgium (Statbel) traite-t-elle ces données?

Depuis 2015, Statbel utilise des scanner data dans le calcul de l'indice des prix à la consommation (IPC). Ces scanner data sont les données scannées aux caisses des magasins (agrégées jusqu'au niveau du produit) et constituent, en plus du webscraping, une source de big data utilisée dans le suivi des prix du panier de l'indice. Les scanner data sont utilisées pour des produits achetés couramment dans les chaînes de supermarchés et remplacent les relevés de prix traditionnels effectués par des enquêteurs.

Statbel reçoit chaque semaine les données des chaînes de supermarchés au moyen de transferts sécurisés. Ces données se composent des informations relatives au produit (codes, descriptions, chiffre d'affaires, etc.) et de la classification interne des chaînes. Dans un premier temps, les produits doivent être classés dans des segments de consommation/groupes de produits prédéfinis, liés à la nomenclature européenne (E)COICOP. Grâce aux données de classification interne et à l'apprentissage automatique, chaque "nouveau" produit (lire "nouveau code de produit") se voit attribuer une classification COICOP, qui peut être retranscrite manuellement si nécessaire.

Les scanner data ne surveillent pas le prix affiché (cf. relevés de prix traditionnels), mais bien le prix moyen réel de la transaction. Cela implique une différence conceptuelle, mais l'impact sur l'évolution des prix semble négligeable. Par ailleurs, l'expérience des scanner data nous apprend qu'il est plus pratique d'utiliser les codes de produit internes (stock keeping units ou SKU) que les codes-barres.

Des indices élémentaires (sans poids) sont d'abord calculés pour toutes les chaînes de supermarchés au niveau du produit (COICOP 6). Les produits représentatifs (= échantillon) sont déterminés par niveau au moyen d'un seuil de chiffre d'affaires, ce qui assure un panier dynamique. Si un produit n'est pas repris dans l'échantillon, des imputations de prix sont calculées. Pour des raisons statistiques et économiques, les indices élémentaires sont calculés à l'aide d'un indice de Jevons (= moyenne géométrique). Bien qu'avec l'utilisation des scanner data, des informations sur le chiffre d'affaires soient également disponibles jusqu'au niveau du produit, nous ne travaillons pas avec des indices pondérés au niveau élémentaire en raison du chain drift. Cependant, le chiffre d'affaires (par segment de consommation et par chaîne) est ensuite utilisé pour agréger les indices aux niveaux supérieurs.

Le marché des produits est loin d'être statique, et les produits connaissent régulièrement des modifications d'emballage, de contenu, de code-barre, etc. Étant donné qu'il faut suivre un même produit pour mesurer l'évolution des prix, il est important de surveiller ces modifications. Lorsqu'une telle modification est repérée dans les scanner data, les codes internes du "nouveau" et de l'"ancien" produit sont reliés. La différence de contenu est prise en compte, de sorte que l'évolution du prix puisse se poursuivre. Par ailleurs, les relances de produits sont liées afin de tenir compte des modifications de prix "cachées".

Les produits saisonniers constituent une catégorie spécifique, étant donné qu'ils sont parfois indisponibles pendant certains mois. C'est pourquoi on utilise pour ces produits la méthode de pondération saisonnière au niveau de la classe au lieu d'un panier dynamique.

Les scanner data et les relevés de prix classiques sont combinés à l'aide d'un modèle de stratification. Il convient également d'établir une distinction entre les supermarchés et les discounters, d'une part, et les magasins spécialisés, d'autres part. Ceux-ci sont alors agrégés sur la base des poids obtenus lors de l'enquête bisannuelle sur le budget des ménages. Les indices mensuels sont reliés à l'aide d'un indice en chaîne. On obtient ainsi un indice à long terme.

L'utilisation des scanner data assure le calcul d'un indice plus représentatif grâce à la disponibilité des données relatives au chiffre d'affaires et à la possibilité de travailler avec des échantillons plus grands.

SOMMAIRE

1. INTRODUCTION

L'indice des prix à la consommation (IPC) est une statistique mensuelle établie par la DG Statistique – Statistics Belgium du SPF Economie (Statbel). Il s'agit d'un indicateur économique qui mesure l'évolution des prix des dépenses de consommation des consommateurs belges. Il est le principal outil de mesure de l'inflation. En Belgique, l'IPC sert de base directe, via l'indice santé et l'indice lissé, à l'indexation des pensions, des allocations sociales, des barèmes fiscaux, des loyers et de certains salaires et traitements.

L'IPC est calculé sur la base d'un panier de biens et de services achetés par les ménages et considérés comme représentatifs de leur comportement de consommation. Étant donné que l'offre de biens et services ne cesse d'évoluer, l'échantillon des prix relevés est également régulièrement actualisé. Actuellement, des prix de biens et de services font l'objet d'un suivi pour 229 catégories de produits.

Ce suivi s'effectue à partir de différentes sources de données. Ainsi, des prix sont relevés par des enquêteurs qui visitent des magasins répartis à travers le pays. La collecte de données pour l'enquête sur les loyers s'effectue soit en format papier, soit en ligne. Les prix présentant les poids les plus importants sont toutefois collectés de manière centralisée vie des sites internet, des catalogues, par téléphone ou via des fichiers obtenus auprès des régulateurs ou d'entreprises privées. Plus récemment, davantage de sources de big data ont également été intégrées au calcul de l'indice des prix à la consommation, à savoir les scanner data des chaînes de supermarchés et les données issues du webscraping.

Outre l'indice national des prix à la consommation (IPC), Statbel calcule également l'indice européen des prix à la consommation harmonisé (IPCH). L'IPCH permet de comparer les taux d'inflation des États membres de l'Union européenne. A cet effet, l'optique des dépenses et les méthodes appliquées sont coordonnées et définies dans la réglementation européenne. Les résultats de l'IPC et de l'IPCH ne sont toutefois pas identiques, en raison principalement de différences de pondération et de composition du panier de biens et de services sur lequel se basent ces indices.

Cet article donne un aperçu de l'utilisation des scanner data, une source de données utilisée dans le calcul de l'IPC depuis 2015 et de l'IPCH depuis 2016. L'ensemble du processus de production est décrit étape par étape :

- quelles sont les scanner data et quelle est leur structure?
- classification des scanner data: comment les produits peuvent-ils être classés dans la bonne hiérarchie des catégories de produits (la nomenclature COICOP[2]) de l'IPC ?
- comment les indices sont-ils calculés à partir de ces données?
- malentendus sur l'utilisation de ces données dans le calcul de l'IPC;
- la méthode de calcul est décrite en détail depuis le niveau du produit jusqu'au niveau agrégé;
- le modèle de stratification permet d'intégrer les indices qui en résultent à d'autres prix provenant d'autres sources de données;
- avantages de l'utilisation des scanner data par rapport à la méthode précédente.

[2] Classification of Individual Consumption by Purpose, une classification des dépenses de consommation conçue par les Nations unies (UNSD – United Nations Statistics Division).

2. DÉFINITION ET APPLICATION

2.1. En quoi consistent les scanner data ou les données scannées à la caisse ?

Eurostat définit les scanner data comme suit:

Transaction data obtained from retail chains containing data on turnover, quantities per item code based on transactions for a given period and from which unit value prices can be derived at item code level[3].

On entend donc par scanner data, les données scannées aux caisses des magasins, agrégées jusqu'au niveau du produit. Il ne s'agit pas des tickets de caisse individuels par client. Actuellement, Statbel reçoit, chaque semaine, ces données de vente agrégées (spécifications du produit, chiffre d'affaires et informations sur les prix) au niveau des produits, à savoir par code-barres, des trois plus grandes chaînes de supermarchés.

Les premières scanner data ont été reçues en octobre 2013 avec les données historiques depuis janvier 2012. Après un an de test, les scanner data ont été intégrées par phases à l'indice des prix à la consommation (IPC), avec une intégration étendue pour 70 groupes de produits en janvier 2016 et l'élargissement à 3 groupes de produits supplémentaires en 2017. Les scanner data sont utilisées dans l'indice des prix à la consommation harmonisé (IPCH) belge depuis 2016 pour les mêmes groupes de produits que dans l'IPC.

Statbel est le 5e office national de statistique ayant mis en œuvre l'utilisation des scanner data dans le calcul des indices des prix à la consommation. Il est ainsi un des précurseurs européens de l'application des scanner data. Par ailleurs, les méthodes d'utilisation des scanner data d'Eurostat sont (en partie) basées sur la méthodologie appliquée par Statbel.

L'objectif est en tous cas d'augmenter le nombre de chaînes qui fournissent des scanner data et d'étendre les scanner data à d'autres secteurs que les supermarchés, comme par exemple l'habillement ou l'électronique grand public.

2.2. À quels groupes de produits s'appliquent les scanner data ?

Il s'agit de produits achetés couramment dans les chaînes de supermarchés. Au total, les 73 groupes de produits obtenus sur la base de scanner data couvrent 23% du poids du panier de l'indice (tableau 1).

Tableau 1: Groupes de produits pour lesquels les scanner data sont utilisées depuis janvier 2017

COICOP	Description	Poids 2017
01	Produits alimentaires et boissons non alcoolisées	16.4%
02	Boissons alcoolisées et tabac	2.5%
05.5.2.2	Accessoires divers pour la maison et le jardin	0.3%
05.6.1	Biens d'équipement ménager non durables	1.1%
09.3.4.2	Produits pour animaux de compagnie	0.7%
09.5.4.1	Produits de papier	0.1%
09.5.4.9	Matériel pour écrire et dessiner	0.2%
12.1.3	Produits pour soins corporels	1.7%
	Total	23.0%

La nomenclature COICOP est structurée hiérarchiquement, ce qui aboutit au 73 groupes mentionnés ci-dessus. Elle répartit les dépenses de consommation totales (niveau 1) en 12 groupes principaux (niveau 2), qui comportent également différents sous-groupes à deux niveaux inférieurs (niveaux 3 et 4). Au niveau européen, elle fait l'objet d'une harmonisation plus poussée jusqu'au cinquième niveau. Ce niveau le plus bas de l'ECOICOP correspond également au niveau de publication de l'IPC et de l'IPCH. On trouvera en annexe un aperçu complet de tous les groupes de produits pour lesquels les données du scanning sont utilisées. Statbel utilise le 6e niveau de la COICOP, les segments de consommation. Outre l'IPC, la nomenclature

[3] Ce que l'on peut traduire par: données relatives aux transactions obtenues auprès de magasins, qui contiennent des données sur le chiffre d'affaires, les quantités par produit, sur la base des transactions d'une période déterminée, et qui permettent de calculer les prix unitaires au niveau du code du produit.

européenne COICOP (ECOICOP) est également utilisée pour l'IPCH, les comptes nationaux et l'enquête sur le budget des ménages.

Chaque segment de consommation (niveau 6) reçoit ensuite un facteur de pondération. Le facteur de pondération du niveau supérieur étant toujours égal à celui des niveaux sous-jacents. Les 73 groupes de produits pour lesquels des scanner data sont utilisées font dès lors référence au nombre de groupes de produits au plus bas niveau de l'ECOICOP.

Le tableau 2 présente un exemple de ventilation des dépenses de consommation. L'un des 12 groupes principaux s'intitule « Produits alimentaires et boissons non alcoolisées ». Vient ensuite le niveau plus détaillé « Produits alimentaires », dont l'une des catégories s'intitule « Pain et céréales ». Cette catégorie est ensuite subdivisée en 8 groupes de produits.

Tableau 2: Exemple de classification des dépenses de consommation

COICOP	Dénomination	Niveau
0	Dépenses totales	1
01	Produits alimentaires et boissons non alcoolisées	2
01.1	Produits alimentaires	3
01.1.1	Pain et céréales	4
01.1.1.1	Riz	5
01.1.1.2	Farines et autres céréales	5
01.1.1.3	Pain	5
01.1.1.4	Autres produits de boulangerie	5
01.1.1.5	Pizza et quiche	5
01.1.1.6	Pâtes alimentaires et couscous	5
01.1.1.7	Céréales du petit déjeuner	5
01.1.1.8	Autres produits à base de céréales	5

Outre les scanner data, des relevés de prix classiques des produits alimentaires sont encore effectués dans un certain nombre de magasins spécialisés (p. ex. les boulangeries et les boucheries) et de discounters. Le webscraping, l'extraction automatique des données des pages web ("scraping"), est également utilisé dans l'IPCH. Pour le calcul final de l'indice, ces différentes sources de données sont combinées sur la base d'informations relatives au chiffre d'affaires et aux dépenses. Cette méthode garantit une mesure représentative de l'évolution des prix. La combinaison de ces différentes sources de données a par contre comme conséquence que le poids effectif des scanner data s'élève à environ 18-19 %.

2.3. À quoi ressemblent les scanner data?

Trois chaînes de supermarchés transmettent chaque semaine à Statbel les données de la semaine précédente via SFTP (SSH File Transfer Protocol). Les données sont divisées en deux parties. La première contient toutes les informations sur le produit, la deuxième partie repend la classification interne de la chaîne. Les tableaux ci-dessous donnent un exemple fictif des deux datasets.

Les informations sur les produits contiennent:

- une indication de la période à laquelle les données se réfèrent;
- plusieurs codes de produit, des données sur le chiffre d'affaires et les quantités;
- des descriptions détaillées du produit;
- un lien vers le dataset reprenant la classification interne.

Tableau 3: Exemple d'informations sur les produits

Variable	Description	Exemple
DT_STRT	Date de début semaine	2/fév/15
DT_STOP	Date de fin semaine	8/fév/15
CD_PROD_CLASS1	Classification interne level 1	D
CD_PROD_CLASS2	Classification interne level 2	E
CD_PROD_CLASS3	Classification interne level 3	I
CD_PROD_CLASS4	Classification interne level 4	K
NR_ITRL	Code produit interne – 1	8523
NR_ART	Code produit interne – 2	1568
NR_EAN	Code GTIN du produit	5449000000286
TX_BRAND_NL	Description marque – néerlandais	Coca-Cola
TX_BRAND_FR	Description marque – français	Coca-Cola
TX_TYPE_NL	Description type de produit - néerlandais	2L
TX_TYPE_FR	Description type de produit - français	2L
TX_INFO_NL	Description info produit - néerlandais	Regular (PET)
TX_INFO_FR	Description info produit - français	Regular (PET)
MS_VAT_RT	Taux de TVA	6
MS_TRNOVR	Chiffre d'affaires	10000
MS_SALES_UNIT	Quantité vendue	4000
CD_TYPE	Vendu à l'unité ou au poids	Units
MS_PKGG	Valeur emballage	2
MS_PKGG_DESCR	Description emballage (litre, kilo, pièces,...)	L
MS_ALC_RT	Pourcentage d'alcool (%)	0
MS_AV_PRC	Prix moyen (MS_TRNOVR/MS_SALES_UNIT)	2,5

Tableau 4: Exemple de classification interne

Variable	Description	Exemple
CD_PROD_CLASS1	Classification interne level 1	D
TX_PROD_CLASS1	Description classification level 1	Food
CD_PROD_CLASS2	Classification interne level 2	E
TX_PROD_CLASS2	Description classification level 2	Drinks
CD_PROD_CLASS3	Classification interne level 3	I
TX_PROD_CLASS3	Description classification level 3	Lemonades
CD_PROD_CLASS4	Classification interne level 4	K
TX_PROD_CLASS4	Description classification level 4	Regular Cola

3. MÉTHODOLOGIE

En résumé, Statbel utilise un panier dynamique (sauf pour les produits saisonniers, voir la section 3.5 Produits saisonniers) avec un indice de Jevons en chaîne pour traiter les scanner data afin d'obtenir des indices. Cet échantillon dynamique est déterminé sur la base du chiffre d'affaires des différents produits individuels pendant deux mois consécutifs.

Un seuil est ensuite utilisé pour déterminer si un produit est inclus ou non dans l'échantillon. Des imputations de prix sont calculées pour les produits non inclus dans l'échantillon. Un produit individuel est déterminé sur la base du code interne plutôt que des codes-barres.

Les relances de produits sont liées afin de tenir compte des modifications de prix "cachées". Si nécessaire, une correction de quantité est effectuée afin de permettre une comparaison entre l'ancien et le nouveau produit. Voici une représentation schématique de ce processus:

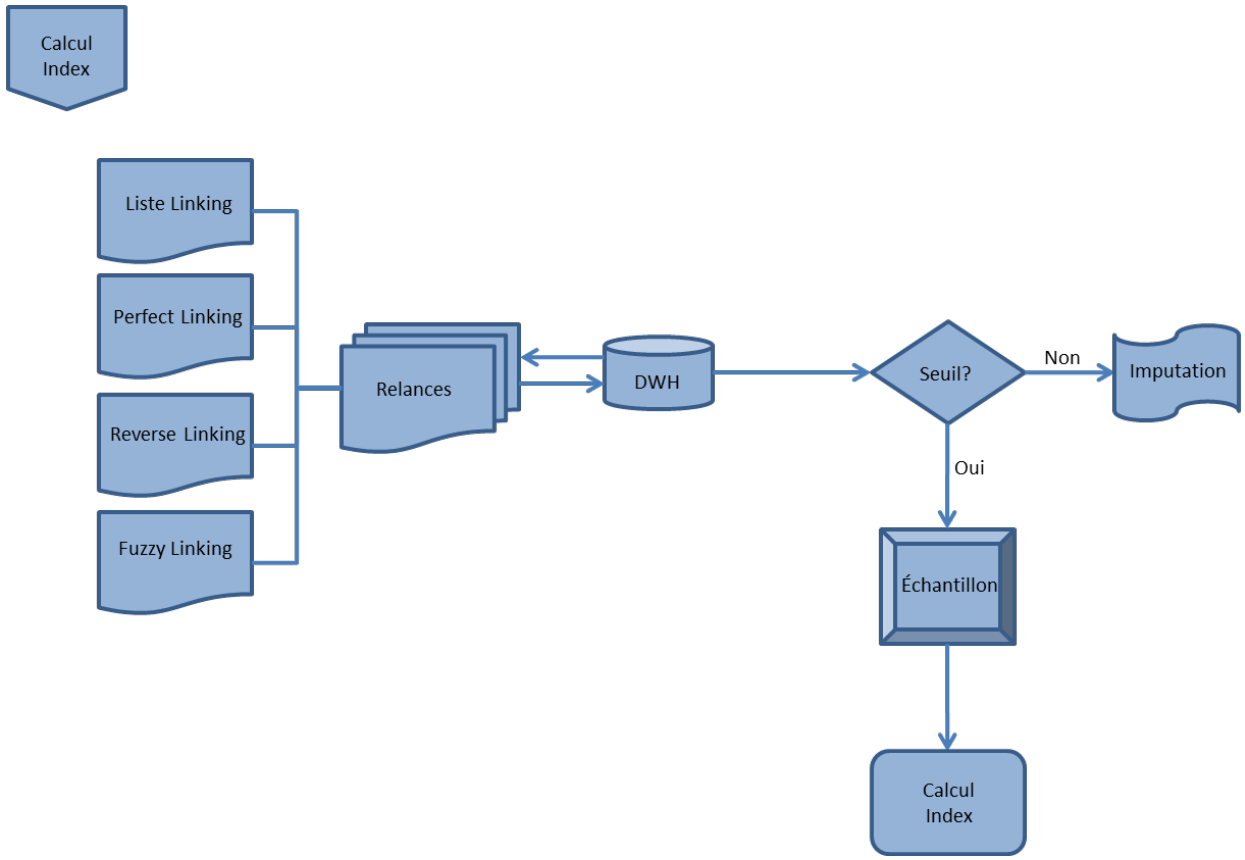


Figure 1: Représentation schématique du calcul de l'indice

Un indice est calculé de cette façon pour chaque chaîne de supermarchés. Ces indices sont ensuite combinés à d'autres données (notamment les relevés de prix classiques) au moyen d'un modèle de stratification.

Les sections suivantes donnent des explications plus détaillées sur chacun des points mentionnés ci-dessus. Avant de pouvoir calculer les indices, les produits doivent être regroupés en groupes de produits/segments de consommation. Ces groupes de produits doivent être liés à la catégorie ECOICOP appropriée.

3.1. Classification

Le travail de classification des scanner data dans la nomenclature COICOP se compose d'une phase de démarrage et d'un traitement récurrent. Durant la phase de démarrage, la classification interne du supermarché est reliée le mieux possible à la classification ECOICOP et différents segments de consommation sont créés au niveau COICOP 6. Un contrôle des nouveaux produits est effectué lors du traitement hebdomadaire.

Le processus de classification est schématisé ci-dessous:

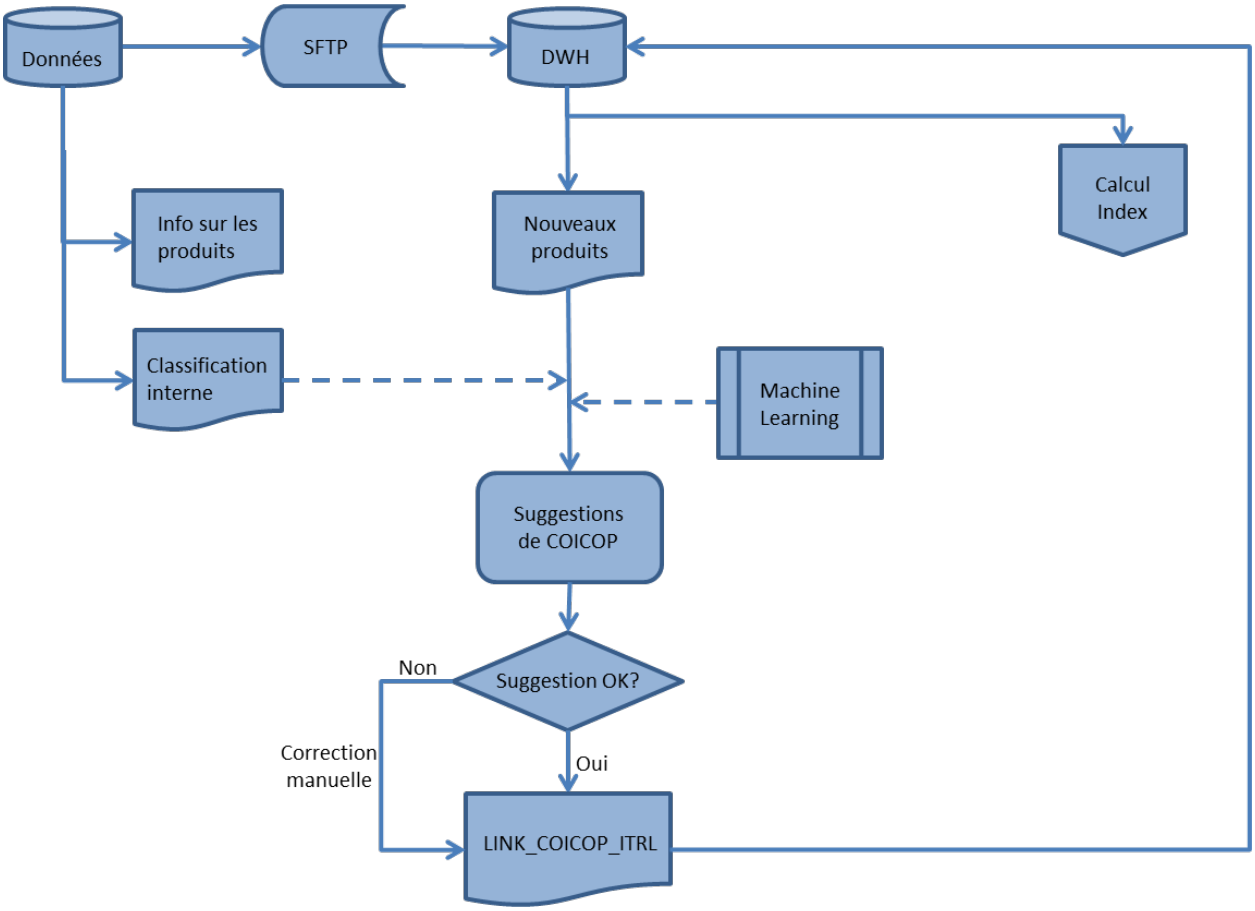


Figure 2: Représentation schématique de la classification

3.1.1. Scanner data et classification COICOP: phase de démarrage

La première étape consiste à relier la classification interne des chaînes de supermarchés à la ECOICOP 5 (pour chaque chaîne séparément). Des subdivisions sont ensuite effectuées par chaîne au niveau de la COICOP 6. Cette opération est réalisée par segment de consommation. Par exemple, la catégorie ECOICOP des boissons rafraîchissantes inclut le coca, la limonade, le thé glacé, etc. Le but n'est pas de créer les mêmes segments pour chaque chaîne, mais, afin de pouvoir comparer les évolutions de prix, on essaye d'obtenir la plus grande concordance possible entre les chaînes. Bien que les scanner data doivent être combinées ultérieurement avec les relevés de prix classiques, le but n'est pas de copier les segments de consommation de la méthode classique. Étant donné que la méthode classique est aussi basée sur un échantillon, limiter les scanner data aux mêmes segments aurait pour conséquence de ne pas prendre en compte beaucoup d'informations et de produits. Par exemple, pour les relevés de prix classiques, on travaille avec environ 173 groupes de consommation par chaîne dans la COICOP 01 (Produits alimentaires et boissons non alcoolisées), alors qu'environ 450 groupes sont créés avec les scanner data.

Chaque chaîne compte environ 3.500 classes internes. Les segments de consommation reposent parfois sur une classification interne de la chaîne, ce qui permet d'affecter l'ensemble de cette classe interne au segment de consommation. Généralement, il s'agit toutefois d'une combinaison de différentes classes internes. Par exemple, la marque de distributeur d'une chaîne est parfois considérée comme une classe interne distincte. Les différentes marques d'un groupe de produits sont parfois même classées séparément. Dans ce cas, les différentes classifications internes sont regroupées dans un segment de consommation de la COICOP 6. De plus, les classifications internes de la chaîne sont parfois trop spécifiques, si bien que la classe ne compte qu'un seul produit. Dans ce cas, plusieurs classes internes seront alors combinées. Lors de l'agrégation ultérieure, des pondérations sont utilisées au niveau de la COICOP 6, ce qui permet d'éviter les segments de consommation ne contenant qu'un seul produit. Si le produit n'était plus disponible et s'il n'existait aucun substitut, l'évolution des prix d'un groupe serait uniquement le résultat d'une imputation, ce qui, selon la réglementation sur l'IPCH, n'est autorisé que pour un mois mais pas plus.

Il est également possible qu'une classification interne soit trop hétérogène pour constituer de bons segments de consommation. Par exemple, la classe interne "vin rouge", qui peut être subdivisée en "vin rouge français","vin rouge italien", etc. ou la classe interne du café peut être scindée en "coffee pads", "café moulu" et "café soluble". Sur la base des données ou par analogie avec les autres chaînes, des sous-classes supplémentaires sont créées au niveau de la COICOP 6. Selon la chaîne, environ 10 à 15 % des produits sont reliés individuellement. En principe, nous évitons de relier au niveau du produit, mais cela s'avère nécessaire en pareil cas.

3.1.2. Scanner data et classification COICOP: travail récurrent

La mise en correspondance des scanner data avec la classification COICOP prend beaucoup de temps au départ. Mais après la phase de démarrage, il suffit de vérifier chaque semaine si les nouveaux produits issus des données du scanning sont reliés au bon groupe au niveau COICOP 6. En cas de lien erroné, le produit est relié individuellement au bon groupe. Selon le nombre de nouveaux produits dans les sets de données, ce processus prend environ un à deux jours par semaine pour les trois chaînes de supermarchés. Une base de données séparée est utilisée pour relier les codes produit internes à un groupe/segment de consommation de la COICOP 6. Si des données se trouvent dans cette base de données, l'autre lien (basé sur la classification interne) est rejeté. En d'autres termes: le lien au niveau du produit est prioritaire par rapport au lien utilisant les classes internes.

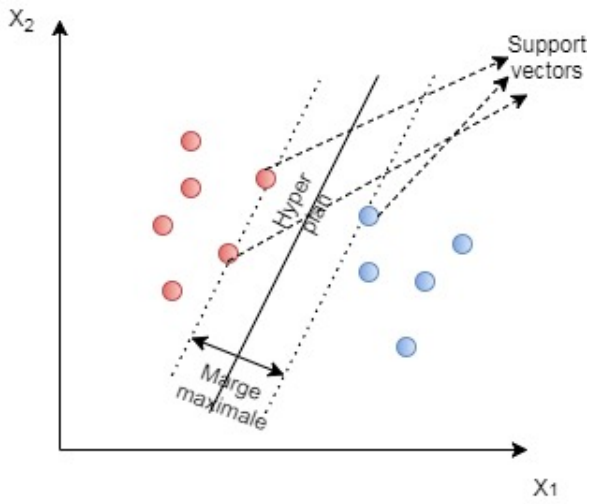
3.1.3. Machine learning

Afin de classer les produits dans le bon segment de consommation de la COICOP, on a recours à "l'apprentissage automatique supervisé" (supervised machine learning - SVM).

L'apprentissage automatique supervisé utilisé par Statbel applique un algorithme de machine à vecteurs de support (support vector machine- SVM). À l'aide d'un dataset d'apprentissage prédéfini, l'algorithme peut construire un modèle pour classifier de nouvelles données en fonction des similitudes entre le dataset d'apprentissage et les nouvelles données. D'un point de vue théorique, un modèle SVM permet la meilleure séparation possible (hyperplan) entre les différentes catégories. En regardant de quel côté de l'hyperplan se trouve le nouvel objet de données, le modèle SVM peut placer l'objet dans la bonne catégorie. La "meilleure séparation possible" signifie que la distance entre l'hyperplan et les objets les plus proches (vecteurs de support) de chaque classe (la marge) est la plus grande que possible.

Le graphique ci-dessous montre de manière schématique comment les observations peuvent être classées en deux classes par SVM linéaire.

Figure 3: Présentation de la classification via l'algorithme de machine à vecteurs de support (support vector machine)



Dans un premier temps, un dataset d'apprentissage est créé en attribuant manuellement les produits à un segment de consommation (partie " supervisée "). L'algorithme de machine à vecteurs de support créera ensuite, sur la base des descriptions de produits et de la classification attribuée, un modèle qui attribuera de nouveaux produits à l'une des catégories. Avant d'être mis en service, l'algorithme est testé sur des données de test qui sont également classées manuellement. Si la marge d'erreur est limitée, l'algorithme peut être appliqué chaque semaine à de nouvelles données. Après cela, la catégorie de produit proposée pour chaque produit est confirmée ou corrigée manuellement. Ces données qui viennent d'être classées sont ajoutées la semaine suivante aux données d'apprentissage, ce qui rend le modèle de plus en plus intelligent et fiable.

Le schéma suivant illustre cette procédure:

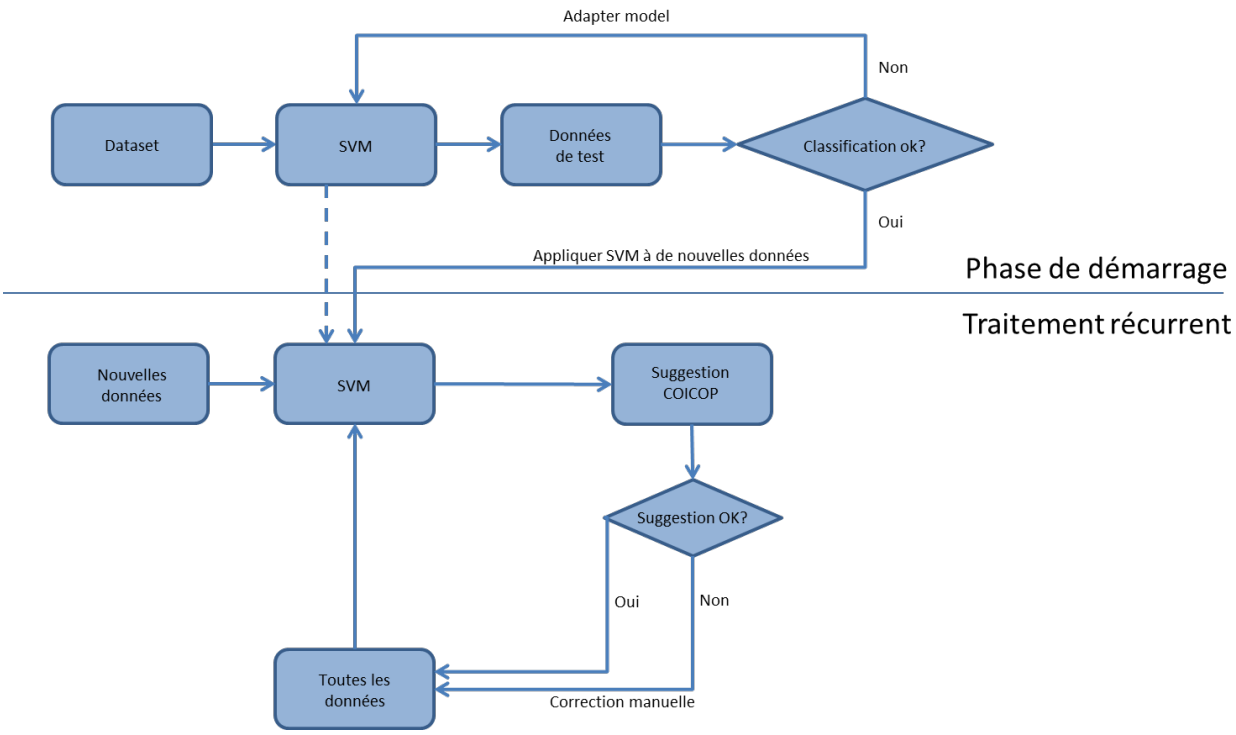


Figure 4: Machine learning pour la classification COICOP

Supposons, par exemple, qu'un supermarché n'ait qu'une seule classe interne pour le café (COICOP 5-groupe 01.2.1.1), mais que, sur la base des produits, quatre sous-classes puissent être établies : pads de café, café torréfié, café moulu et café soluble. L'objectif est de répartir les produits en quatre catégories sur la base de la description (texte). Une partie du dataset sur le café est d'abord classée manuellement selon les quatre catégories (dataset d'apprentissage). L'algorithme établit ensuite un modèle, qui est à son tour testé sur les données de test. S'il est évalué positivement, l'algorithme peut être appliqué "en production" à de nouvelles données. Après cela, ces nouvelles données correctement classées sont utilisées pour réévaluer le modèle pour la classification des données futures.

Il existe également un apprentissage automatique non supervisé, qui ne nécessite aucun dataset prédéfini. L'algorithme détermine lui-même les catégories sur la base des données. L'avantage de cette méthode est qu'il ne faut établir aucun dataset d'apprentissage manuellement. L'inconvénient est que le résultat ne crée probablement pas les catégories comme on l'aurait attendu. Dans l'exemple du café, il se peut qu'un segment de consommation contenant à la fois le café torréfié et les pads de café soit créé (plutôt que deux catégories distinctes). Un autre point négatif peut être que, comme cette méthode ne recherche pas les similitudes entre les descriptions, des segments de consommation différents peuvent être créés entre deux périodes, parce que de nouveaux produits ont été ajoutés.

Ces inconvénients ont poussé Statbel à recourir à l'apprentissage automatique supervisé.

3.2. Concept de prix

3.2.1. Prix unitaires (de valeurs unitaires)

Contrairement aux prix affichés utilisés pour les relevés de prix traditionnels, les scanner data permettent d'observer le prix moyen réel de la transaction. Ce prix est calculé comme le quotient du chiffre d'affaires total et de la quantité vendue d'un produit pendant une période donnée. En Belgique, cette durée est généralement d'une ou deux semaines pour l'IPC et trois semaines pour l'IPCH

Les chiffres d'affaires et les quantités sont agrégés par produit au niveau des chaînes de supermarchés, ce qui permet d'obtenir un prix de transaction moyen (ou prix unitaire) par produit et par chaîne. Les relevés de prix traditionnels permettent également d'obtenir un prix moyen par produit et par chaîne, mais seulement un certain jour du mois. Les scanner data fournissent des informations sur les prix sur une plus longue période de temps. Ainsi, bien qu'il existe une différence conceptuelle entre les prix observés, l'impact sur l'évolution des prix est toutefois négligeable, comme le montrent les graphiques suivants[4]:

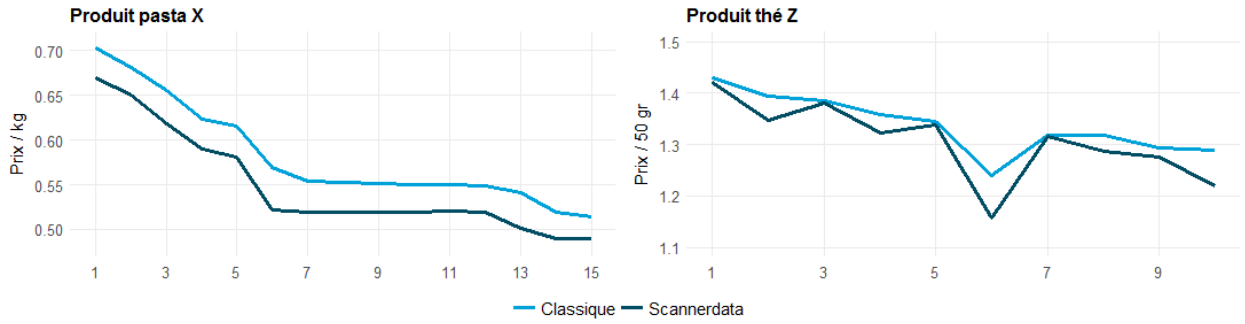


Figure 5: Comparaison de l'évolution des prix entre la méthode classique et les scanner data

Les prix des scanner data sont généralement inférieurs aux prix enregistrés lors des relevés traditionnels. C'est dû à la plus grande quantité de prix captés au moyen des scanner data. Les scanner data incluent également les promotions et les réductions, qui seront moins observées avec la méthode classique, étant donné que la réglementation de l'IPCH impose d'exclure les réductions discriminatoires si aucune information n'est disponible sur le nombre de personnes ayant acheté un produit en promotion.

Les réductions discriminatoires sont des réductions qui ne sont pas accordées à tout le monde (seulement d'application avec une carte de fidélité) ou qui ne sont valables que pendant une journée déterminée. Dans le cas des relevés de prix classiques effectués par les enquêteurs, ces informations ne sont pas disponibles et, par conséquent, conformément à la réglementation de l'IPCH, ces réductions ne sont pas prises en compte. Toutefois, avec les scanner data, ces données sont bel et bien disponibles car le prix moyen comprend les transactions ayant bénéficié d'une réduction. Ces réductions peuvent dès lors être prises en compte, conformément à la réglementation de l'IPCH. Malgré ces prix plus bas, l'évolution des prix est quasi identique.

La différence de nombre de semaines pour le calcul de l'IPCH et de l'IPC est due à la différence de date de publication des deux indices. L'IPC est publié l'avant-dernier jour ouvrable du mois. L'IPCH est publié par Statbel et Eurostat au plus tard deux semaines après la fin du mois. La règlementation de l'IPCH stipule également que le calcul de l'indice doit utiliser les prix de la semaine du 15e jour du mois. Comme l'IPC est publié tôt, il est normalement impossible de satisfaire à cette exigence pour l'IPCH. De plus, Eurostat recommande d'utiliser les prix de trois semaines pour calculer des indices au moyen de données du scanning. La proposition de retarder la date de publication de l'IPC et d'avancer la date de publication de l'IPCH - à l'instar d'autres pays européens - afin que les deux indices soient publiés en même temps et utilisent les mêmes informations sur les prix, a été rejetée par la Commission de l'indice[5] en raison des implications sur les mécanismes d'indexation en vigueur en

[4] Les résultats expérimentaux sont basés sur plusieurs périodes différentes. Pour la généralité, celles-ci sont toujours numérotées en commençant par 1. Chaque période correspond à un mois.

[5] La Commission de l'indice est composée, de manière paritaire, des organisations patronales et syndicales et de représentants du monde académique. La Commission bénéficie du soutien des statisticiens de Statbel. Cette Commission conseille le ministre de l'Economie sur l'ensemble des questions relatives à l'indice des prix à la consommation et émet chaque mois un avis sur l'indice calculé par Statbel. Elle rend également son avis sur l'actualisation annuelle au ministre de l'Economie.

Belgique.La différence d'agrégation des prix sur des périodes plus courtes dans l'IPC n'a pas d'impact à long terme sur l'évolution de l'indice. A court terme, on observe toutefois des différences, comme nous l'expliquons dans le paragraphe suivant.

3.2.2. Agrégation des prix sur des périodes différentes

Les prix au niveau des produits individuels sont donc calculés en agrégeant le chiffre d'affaires et les quantités vendues sur la période sur laquelle le calcul de l'indice est basé, puis en prenant le quotient.

Comme décrit ci-dessus, les calculs de l'IPCH et l'IPC utilisent des semaines différentes. Le chapitre 3.3 explique comment le calcul de l'indice est effectué avec les scanner data. Cependant, nous abordons déjà la manière dont le prix unitaire est calculé à l'aide des scanner data et de son effet éventuel sur l'évolution mesurée des prix.

En effet, inclure plusieurs semaines dans le calcul de l'IPCH a pour conséquence que les réductions et les promotions - qui durent généralement une semaine - ont un impact moins important sur le prix unitaire parce que leur effet est lissé sur plusieurs semaines. Il en résulte des indices plus stables en glissement mensuel, comme le montre le graphique suivant pour le COICOP 12.1.3. L'évolution des prix à long terme est toutefois identique.

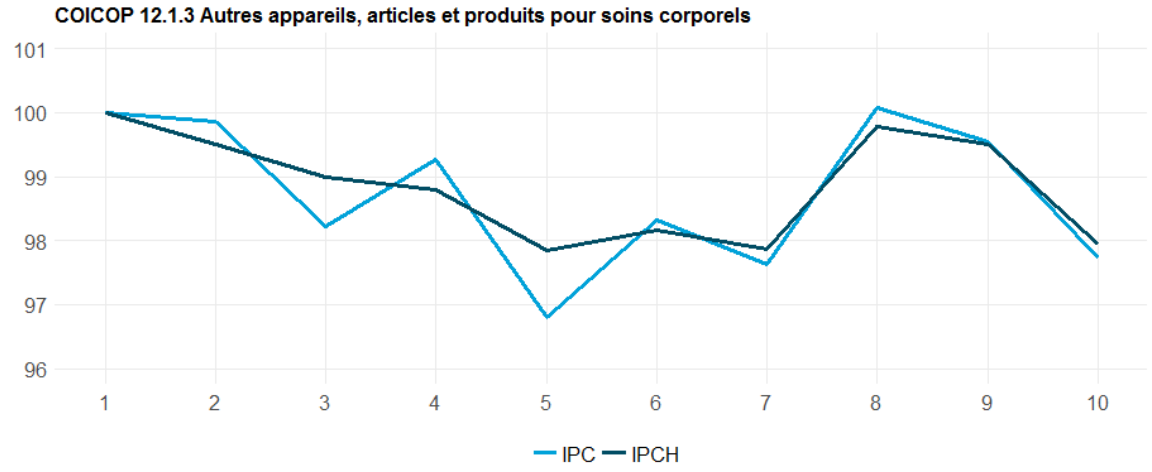


Figure 6: indices (IPC et IPCH) pour le groupe COICOP 12.1.3

Aux niveaux agrégés, la différence entre l'IPC et l'IPCH est toutefois plus faible. Par exemple, la différence pour le COICOP 01 global (Figure 7) est négligeable étant donné que les promotions aux niveaux inférieurs de la COICOP s'annulent d'un mois à l'autre. La fin d'une promotion crée un effet à la hausse tandis qu'une nouvelle promotion dans un autre groupe de produits provoque un mouvement inverse, les deux effets s'annulant ainsi dans l'agrégation.

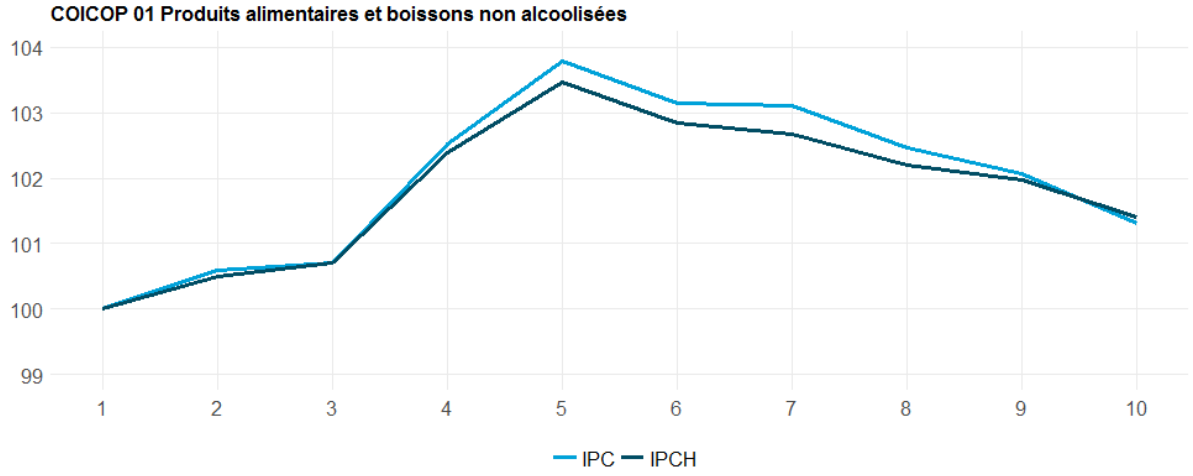


Figure 7: indices agrégés (IPC et IPCH) pour le groupe COICOP 01

Le calcul sur la base de périodes de temps différentes débouche sur la même évolution des prix à long terme. La méthodologie utilisée garantit donc la cohérence, quel que soit le nombre de semaines utilisé.

Il convient également de s'attarder sur le choix de calculer un prix unitaire par produit sur une période de plusieurs semaines. Ce calcul reflète le prix de transaction réel auquel un produit a été acheté, mais diffère considérablement de la façon dont les prix sont collectés manuellement par les enquêteurs. Avec la collecte manuelle des prix, les informations sur les volumes de vente (ou le chiffre d'affaires) manquent et on utilise donc seulement les prix non pondérés d'un produit individuel.

Si l'on tente de reproduire cette méthode avec des scanner data en utilisant une moyenne non pondérée des prix quotidiens sur une période de trois semaines par rapport à un prix unitaire sur la même période, on obtient à nouveau une évolution de prix globalement identique (figure 8). L'indice basé sur les prix journaliers est bien sûr plus stable que l'indice qui agrège les prix sur une période plus longue car l'effet des réductions est ici aussi lissé.

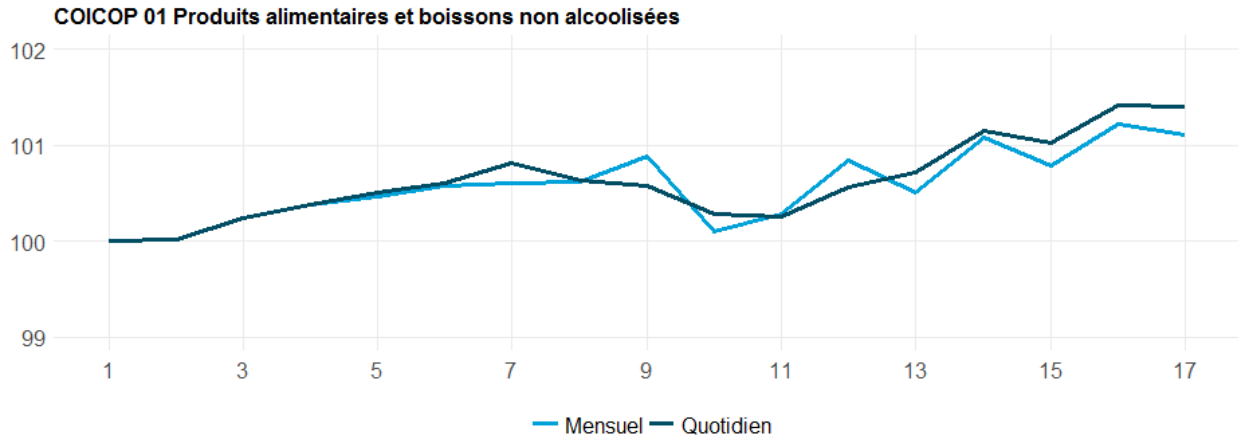


Figure 8: Évolution des prix unitaires quotidiens et mensuels

Il existe d'ailleurs également une forte corrélation entre les prix quotidiens et hebdomadaires. Dans 83% des cas, elle est supérieure à 99% (voir figure 9).

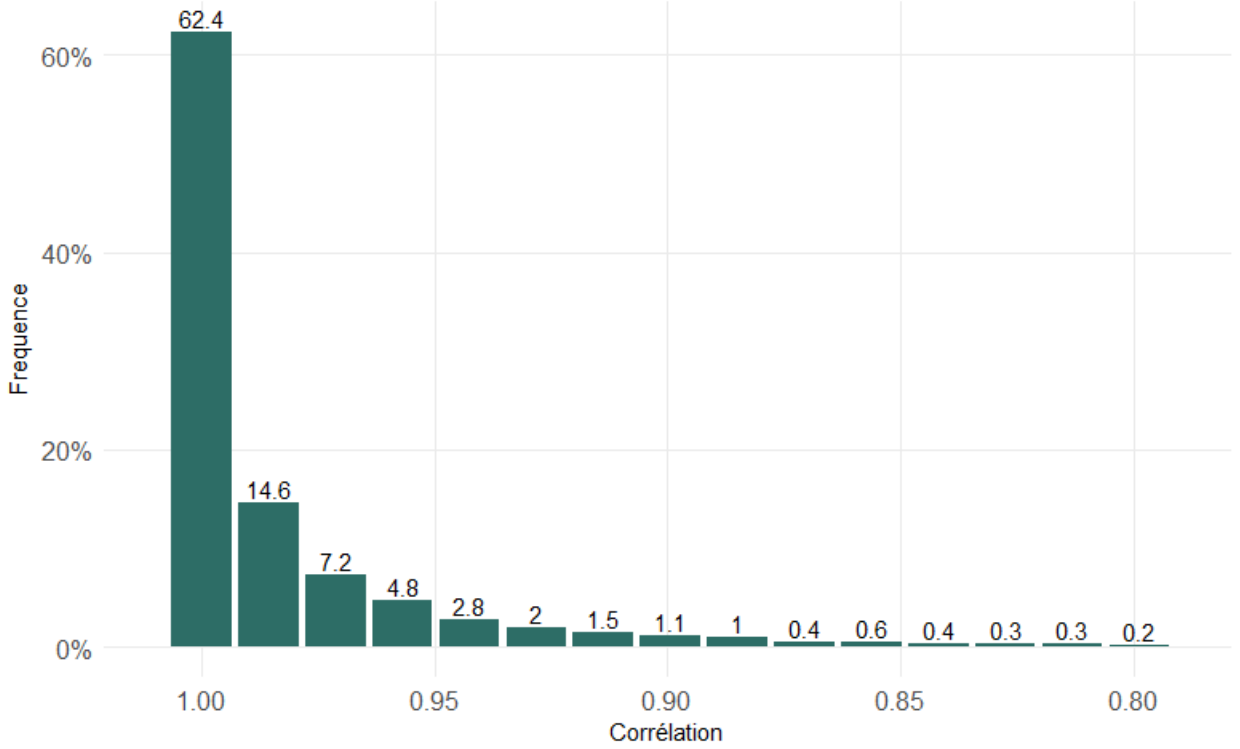


Figure 9: corrélation entre les prix par jour et les prix par semaine

3.2.3. Agrégation des prix entre différents magasins

Avec les scanner data, les prix sont agrégés au niveau d'une chaîne de supermarchés. Cela simplifie la méthode de calcul. Néanmoins, s'il y a une différence de "niveau de service" cela peut potentiellement engendrer un biais dans l'évolution mesurée des prix. Il peut s'agir d'une différence entre les magasins d'une même marque de chaîne (par exemple un même type de magasin dans différentes communes), ou d'une différence entre les marques d'une chaîne (par exemple les petits magasins de quartier et les grands supermarchés). En effet, l'évolution des prix peut différer d'un magasin ou d'un segment à l'autre et le passage des clients à d'autres magasins au sein d'une même chaîne de supermarchés devrait théoriquement être neutralisé par la ventilation et le calcul d'un indice par magasin ou par segment.

Dans la pratique, un calcul effectué au niveau du point de vente individuel, pour être ensuite agrégé au niveau de la chaîne, produit des résultats similaires à un calcul effectué directement au niveau de la chaîne de magasins. L'agrégation entre les magasins ne pose donc aucun problème.

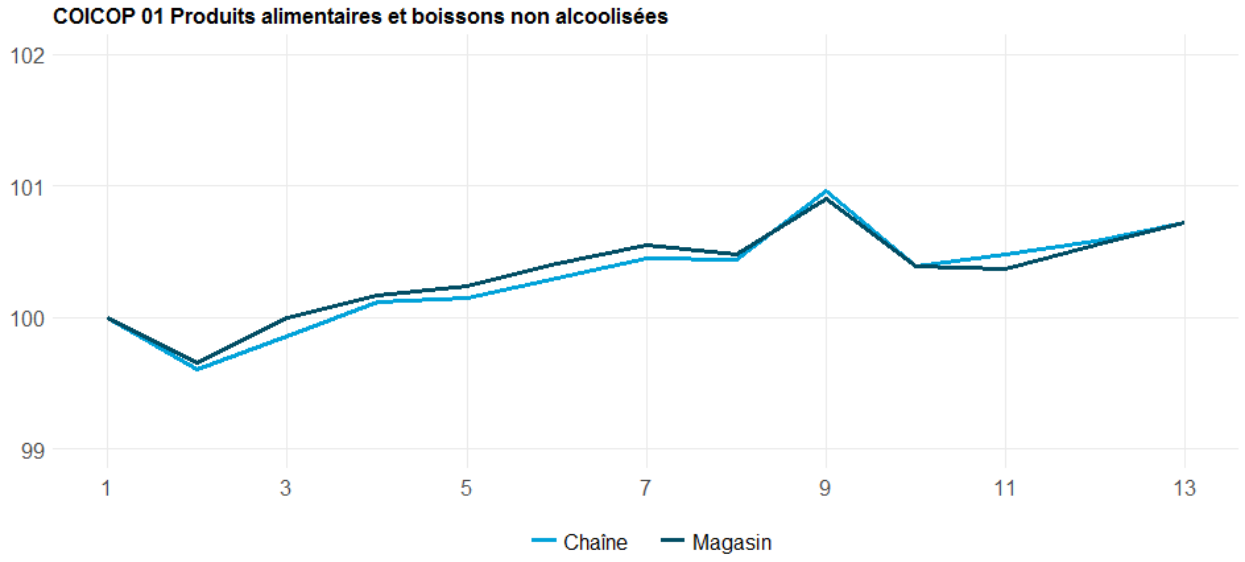


Figure 10: Évolution des prix au niveau de la chaîne et avec agrégation au niveau des magasins

L'agrégation des différents segments de magasins d'une même chaîne n'entraîne pas non plus de biais. La méthode selon laquelle une ventilation est d'abord effectuée par segment puis agrégée par segment n'entraîne pas de biais dans l'évolution des prix mesurée par rapport à la méthode selon laquelle le prix unitaire est calculé directement par produit sur l'ensemble des segments de magasins.

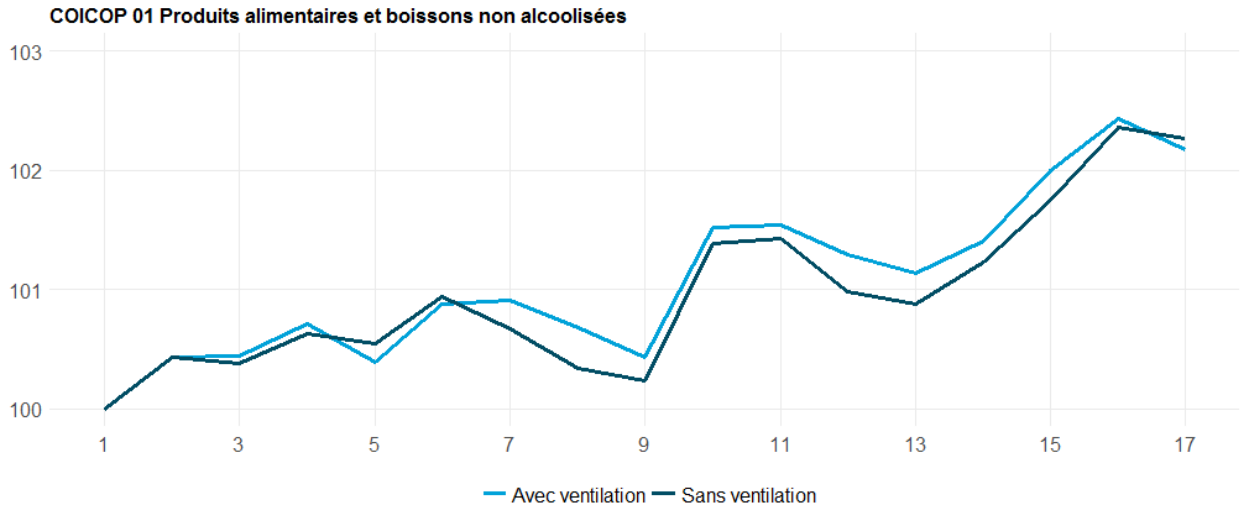


Figure 11: Évolution des prix avec et sans ventilation entre les segments de magasins

3.2.4. Code interne ou code-barres

La mesure de l'inflation a pour objectif de mesurer l'évolution du prix d'un même produit au fil du temps. Si différents produits sont mélangés, on mesure plutôt l'évolution moyenne des dépenses et pas l'évolution des prix. Afin de relier les mêmes produits mois après mois, on utilise au niveau international le GTIN (Global Trade Item Number), mieux connu sous le nom de code-barres. Il s'agit d'un code unique attribué à un produit particulier. Cependant, l'expérience des scanner data de Statbel a montré qu'il est plus pratique d'utiliser les codes internes des chaînes de supermarchés (stock keeping units ou SKU). Ils sont plus stables et uniques pour le calcul de l'indice des prix. Ils sont également utilisés dans d'autres pays, comme en Suisse et en Australie. Cependant, les codes GTIN sont utilisés pour vérifier la cohérence entre les différents supermarchés au niveau de la liaison avec la classification COICOP. Les exemples suivants montrent pourquoi l'on préfère le code interne au code-barres.
Le tableau ci-dessous montre l'exemple d'un produit qui a été vendu pendant plusieurs semaines avec différents codes-barres mais qui a été lié au même code interne.

Tableau 5: Exemple de produit avec différents codes-barres

Semaine	Code interne	Description du produit	Unité	Unités vendues	Chiffre d'affaires	Prix	EAN
3713	12345	Marque x - 40 pièces	0,375	380	2755	7,25	#8000565755675
3813	12345	Marque x - 40 pièces	0,375	561	3540	6,31	#8000565755675
3913	12345	Marque x - 40 pièces	0,375	1289	7657	5,94	#8000565755675
4013	12345	Marque x - 40 pièces	0,375	763	4288	5,62	#8000565755675
4113	12345	Marque x - 40 pièces	0,375	1128	6757	5,99	#8000565755675#8000508890089
4213	12345	Marque x - 40 pièces	0,375	912	5591	6,13	#8000565755675#8000508890089
4313	12345	Marque x - 40 pièces	0,375	621	4229	6,81	#8000565755675#8000508890089
4413	12345	Marque x - 40 pièces	0,375	848	5080	5,99	#8000565755675#8000508890089
4513	12345	Marque x - 40 pièces	0,375	2120	12699	5,99	#8000565755675#8000508890089
4613	12345	Marque x - 40 pièces	0,375	6728	44270	6,58	#8000565755675#8000508890089

Le code interne prouve qu'il s'agit du même produit. Bien que différents codes-barres soient utilisés, il s'agit toujours du même produit, vendu au même prix. Le tableau suivant le montre aussi.

Tableau 6: Exemple de produit avec différents codes-barres pendant la même semaine

EAN	Code interne	Semaine	Description du produit	Unité	Unités vendues	Chiffre d'affaires	Prix
8000565755675	12345	4113	Marque x - 40 pièces	0,375	410	2455,9	5,99
8000508890089	12345	4113	Marque x - 40 pièces	0,375	718	4300,82	5,99

De plus, un code-barres est parfois réutilisé pour un produit complètement différent. Les multipacks reçoivent également souvent un code-barres différent pour accélérer le traitement à la caisse, mais ce code-barres est parfois lié au même code produit interne. Ce code interne peut également être utilisé pour consulter les produits sur le site internet des chaînes de supermarchés. Cela peut permettre de relier les anciens et les nouveaux produits en cas de remplacement. Les codes internes permettent également de calculer un indice pour les produits frais (par exemple, la viande, les fruits et les légumes). Cependant, ces produits sont vendus en quantités différentes (avec des poids différents), chaque poids ayant un code-barre distinct. Les codes-barres utilisés pour ces produits ne sont d'ailleurs pas des codes GTIN officiels, mais des codes-barres générés en interne ou codes PLU (price look-up), qui sont encodés à la caisse. En agrégeant les différentes quantités sur la base du code interne, il est possible de déterminer le prix moyen au kilo.

3.3. Échantillonnage et agrégation

Cette section présente le calcul de l'indice au niveau le plus bas au sein d'une chaîne de supermarchés.

3.3.1. Calcul d'indices de prix élémentaires

Au départ, des indices élémentaires sont calculés pour chaque chaîne séparément. Les indices élémentaires sont ceux pour lesquels aucune pondération n'est utilisée. Par contre, l'indice qui en résulte est quant à lui pondéré. Les indices élémentaires des scanner data se situent au niveau 6 de la COICOP et sont calculés sur la base des données de prix individuelles, au niveau du produit (niveau des SKU).
Comme tous les produits au sein d'un groupe ont le même poids, il est indispensable de composer un échantillon qui ne contient que des produits représentatifs. Pour ce faire, on a recours à un seuil (treshold). Sans seuil, l'indice serait influencé par des produits qui sont peu achetés. On constate en effet (voir plus loin) qu'au sein des groupes, un petit nombre de produits affichent une part de marché importante. Un produit n'est repris dans l'échantillon que si sa part de marché moyenne sur deux mois dépasse un certain seuil, qui dépend du nombre de produits par groupe.

Un produit est repris dans l'échantillon quand

$$\frac{s_m + s_{m-1}}{2} > \frac{1}{n * \lambda} \qquad \textit{Équ. 1}$$

avec λ = 1,25

n = nombre de produits par groupe

s_m = part de marché de chaque produit pendant le mois m

s_{m-1} = part de marché de chaque produit pendant le mois m-1

L'expérience montre qu'un nombre relativement faible de produits représente une part importante du chiffre d'affaires d'un segment de consommation de la COICOP 6. En moyenne, entre 40 et 45 % des produits représentent environ 80 % des parts de marché. Le tableau suivant présente quelques exemples pour le groupe ECOICOP 5 des yaourts (1.1.4.4):

Tableau 7: Part de marché COICOP 1.1.4.4

COICOP	Description	% produits	Part de marché totale
01.1.4.4.01	Yaourt maigre (nature)	46.67%	81.72%
01.1.4.4.02	Yaourt entier (nature)	45.00%	79.05%
01.1.4.4.03	Yaourt maigre aux fruits	37.50%	71.56%
01.1.4.4.04	Yaourt entier aux fruits	37.74%	81.96%
01.1.4.4.05	Yaourt fonctionnel	72.22%	92.70%
01.1.4.4.06	Yaourt pour enfants	80.00%	85.32%
01.1.4.4.07	Yaourt à boire	44.00%	70.77%

Dans le groupe ECOICOP 5 des boissons rafraîchissantes, 80% des parts de marché sont déterminés par moins de 30% des produits.

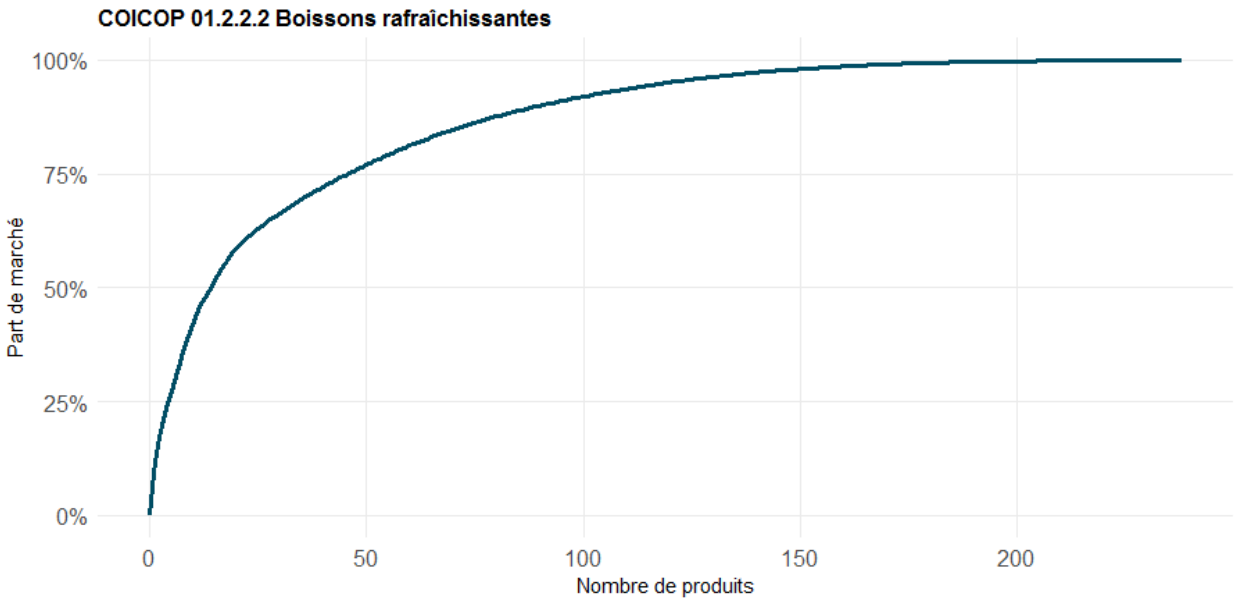


Figure 12: part de marché par rapport au nombre de produits dans la COICOP 1.2.2.2

La formule montre que le calcul de l'indice s'effectue au moyen de prix relatifs (soit l'évolution des prix) des produits qui se trouvent pendant deux mois consécutifs dans l'échantillon, sur la base d'un seuil dynamique.

Les prix relatifs sont déterminés au niveau du produit (code interne). La comparaison s'effectue chaque fois sur le prix du même produit d'un mois à l'autre. Afin d'obtenir un indice élémentaire, ces prix relatifs sont agrégés au sein de leur segment de consommation au moyen d'une moyenne géométrique (indice de Jevons), comme c'est le cas depuis 2014 pour les relevés de prix classiques. Ce niveau du groupe, qui n'est pas publié car il est propre à la chaîne concernée, se situe en-dessous du niveau COICOP 5, qui est lui publié.

Tout cela montre que, comme auparavant, l'indice des prix sur la base des scanner data reflète toujours l'évolution des prix d'un produit identique (p.ex. une cannette de limonade de 33 cl de la marque A dans un supermarché X) pendant un mois déterminé par rapport à son prix durant le mois précédent. Dans l'indice, le suivi est toujours réalisé tant pour les produits premier prix que pour les produits de marque distributeur ou encore pour les produits de marque, pour autant qu'ils atteignent des ventes représentatives durant la période considérée et soient donc repris dans l'échantillon.

Cependant, il est erroné de conclure que l'utilisation des scanner data entraîne une baisse de l'indice dans le cas d'une plus forte consommation de produits meilleur marché au détriment de produits plus chers ou engendre une hausse de l'indice dans le cas inverse. L'indice des prix à la consommation reflète l'évolution pure des prix. Les changements dans le comportement d'achat des consommateurs ne peuvent pas avoir et n'auront aucun impact direct sur l'évolution des prix mesurée. Au niveau des produits (niveau des codes-barres), il n'est pas tenu compte des volumes vendus, ni des chiffres d'affaires lors de l'agrégation des différents produits pour obtenir l'indice d'un groupe de produits. L'agrégat est une moyenne géométrique non pondérée. Si l'agrégation des prix se basait sur les chiffres d'affaires, on obtiendrait alors un indice de valeur unitaire qui ne mesure pas l'évolution pure des prix, mais bien une combinaison des fluctuations des prix et des quantités vendues.

Les indices élémentaires obtenus pour chaque chaîne sont agrégés au moyen des chiffres d'affaires des segments de consommation concernés atteints l'année précédente. L'agrégation est pondérée sur la base des poids de chaque groupe élémentaire. Ces poids sont adaptés chaque année et sont dérivés du chiffre d'affaires de chaque groupe élémentaire réalisé l'année précédente. Les poids des indices élémentaires restent inchangés pendant l'année. Les changements de produits représentatifs ont lieu au niveau du produit, un sous-niveau des indices élémentaires. En utilisant les scanner data, le choix des produits représentatifs devient objectif.

Pour l'étape d'agrégation suivante, les indices sont agrégés par chaîne en fonction du chiffre d'affaires des chaînes respectives. Pour obtenir les poids, qui sont une mesure du chiffre d'affaires, on utilise les comptes annuels ou les déclarations à la TVA des différentes chaînes.

Les prix des produits temporairement manquants sont imputés. Cette imputation est nécessaire pour pouvoir tenir compte de l'évolution entre le dernier mois au cours duquel le produit était disponible et le mois au cours duquel le produit est remis à disposition. L'imputation des prix manquants s'effectue sur la base de l'évolution des prix des produits analogues. L'absence

du produit n'a pas d'impact sur l'indice mais on tient compte de l'évolution des prix "manquante" quand le produit réapparaît dans l'échantillon.

Voici deux exemples de la procédure décrite ci-dessus. Le premier exemple porte sur la détermination de l'échantillon. L'imputation est examinée dans le deuxième exemple.

Exemple 1: détermination de l'échantillon

Supposons les données suivantes: 4 produits différents, le produit 2 n'étant disponible qu'à partir de février et le produit 4 retiré du marché en mars.

Prix	Déc-16	Jan-17	Fév-17	Mars-17	Avr-17
produit 1	6,00	6,30	6,20	6,40	6,00
produit 2			4,50	4,60	4,40
produit 3	5,20	5,40	5,20	5,30	5,10
produit 4	5,70	5,40	5,80		

Les chiffres d'affaires des 4 produits sont présentés dans le tableau suivant:

Chiffre d'affaires	Déc-16	Jan-17	Fév-17	Mars-17	Avr-17
produit 1	170	200	140	120	110
produit 2			90	180	170
produit 3	80	100	60	70	140
produit 4	90	120	70		
Total	340	420	360	370	420

Le produit 2 ne peut pas encore être repris dans l'échantillon en février, étant donné qu'il n'y a aucune évolution de prix par rapport au mois de janvier. À tout moment, n (nombre de produits) est donc égal à 3, même en février, alors que 4 produits sont disponibles. La formule mentionnée porte le seuil de la part de marché moyenne à environ 26,7 %.

Les données ci-dessus permettent de calculer la part de marché pour les périodes s_m et s_{m-1}. Pour le mois de février, le calcul doit être effectué deux fois. La première fois (a) pour le calcul de l'indice jusqu'en février, lorsque le produit 2 ne se trouve pas encore dans l'échantillon. La deuxième fois (b), la part de marché de février est recalculée en mars, étant donné que le produit 2 de mars peut être comparé avec celui de février. Le produit 4 sort toutefois de l'échantillon car il n'est plus disponible en mars.

Part de marché	Déc-16	Jan-17	Fév-17 (a)	Fév-17 (b)	Mars-17	Avr-17
produit 1	50.0%	47,6%	51,9%	48,3%	32,4%	26,2%
produit 2				31,0%	48,6%	40,5%
produit 3	23,5%	23,8%	22,2%	20,7%	18,9%	33,3%
produit 4	26,5%	28,6%	25,9%			

Nous pouvons dès lors déterminer les produits dont la part de marché moyenne est supérieure au seuil de 26,7 % pendant deux mois consécutifs (m et m-1). Ces produits seront repris dans le calcul de l'indice du mois m.

Part de marché moyenne	Déc-16	Jan-17	Fév-17		Mars-17	Avr-17
produit 1		48,8%	49,7%		40,4%	29,3%
produit 2					39,8%	44,6%
produit 3		23,7%	23,0%		19,8%	26,1%
produit 4		27,5%	27,2%			

Cela signifie que le produit 1 se trouve dans l'échantillon pendant toute la période, alors qu'à partir du mois de mars, le produit 4 sera remplacé dans l'échantillon par le produit 2.

L'indice des prix élémentaire au niveau de la COICOP 6 est calculé à l'aide de l'indice de Jevons en chaîne:

Indice	Déc-16	Jan-17	Fév-17	Mars-17	Avr-17
produit 1		1,05	0,98	1,03	0,94
produit 2				1,02	0,96
produit 4		0,95	1,07		
Indice mensuel		1,00	1,03	1,03	0,95
Indice en chaîne	100,00	99,74	102,54	105,33	99,75

Exemple 2: imputation

Comme indiqué précédemment, les prix des produits exclus de l'échantillon sont imputés. Cette imputation s'effectue en multipliant le dernier prix (lorsque le produit faisait encore partie de l'échantillon) par l'indice de Jevons du groupe de la COICOP 6 auquel appartient le produit. De même, lorsqu'un produit a un prix réel, mais est exclu parce que sa part de marché moyenne n'est pas suffisamment importante (seuil), le prix du produit est imputé. Cette imputation est nécessaire car dans le cas contraire, la variation de prix entre le dernier mois au cours duquel le produit se trouve dans l'échantillon et le mois au cours duquel ce produit réapparaît dans l'échantillon ne serait pas prise en compte.

En théorie, l'imputation veille aussi à ce que l'indice réponde tant au test d'identité qu'au test de transitivité. Le test d'identité vérifie que, si les prix d'un même produit de l'échantillon pendant la période en cours t sont égaux aux prix de la période de référence/de base 0, l'indice est égal à 100. L'indice est transitif lorsque le couplage des indices en chaîne mensuels est égal à la comparaison de prix directe entre les périodes 0 et t. Sans imputation, aucun des deux tests ne serait satisfait et l'évolution des prix mesurée ne serait pas correcte.

Prenons à nouveau 4 produits, dont les produits 2 et 4 sont exclus de l'échantillon respectivement en février et mars, bien qu'ils soient toujours vendus. Par souci de simplicité, cet exemple ne tient pas compte du chiffre d'affaires. Les prix de décembre et d'avril sont identiques. Donc, selon le test d'identité, l'indice en chaîne de décembre à avril doit être égal à 100.

Prix	Déc-16	Jan-17	Fév-17	Mars-17	Avr-17
produit 1	6,00	6,30	6,20	6,40	6,00
produit 2	4,60	4,70		4,30	4,60
produit 3	5,20	5,40	5,20	5,30	5,20
produit 4	5,70	5,40	5,80		5,70

Le tableau suivant présente les indices mensuels sans imputation. L'indice en chaîne s'élève à 104,02 en avril, soit un écart de 4,02 % de l'indice 100 que l'on doit obtenir selon le test d'identité.

Indices	Déc-16	Jan-17	Fév-17	Mars-7	Avr-17
produit 1		1,05	0,98	1,03	0,94
produit 2		1,02			1,07
produit 3		1,04	0,96	1,02	0,98
produit 4		0,95	1,07		
Indice mensuel		1,01	1,01	1,03	0,99
Indice en chaîne	100,00	101,36	101,96	104,58	104,02

Tout d'abord, le prix du produit 2 est imputé en février, en multipliant le dernier prix, lorsque le produit 2 faisait partie de l'échantillon (janvier), par l'indice mensuel de Jevons de février.

Prix	Déc-16	Jan-17	Fév-17	Mars-17	Avr-17
produit 1	6,00	6,30	6,20	6,40	6,00
produit 2	4,60	4,70	4,73	4,30	4,60
produit 3	5,20	5,40	5,20	5,30	5,20
produit 4	5,70	5,40	5,80		5,70

Ce prix obtenu pour le produit 2 en février permet désormais de calculer un indice pour ce produit en mars.

Indices	Déc-16	Jan-17	Fév-17	Mars-7	Avr-17
produit 1		1,05	0,98	1,03	0,94
produit 2		1,02	1,01	0,91	1,07
produit 3		1,04	0,96	1,02	0,98
produit 4		0,95	1,07		
Indice mensuel		1,01	1,01	0,99	0,99
Indice en chaîne	100,00	101,36	101,96	100,47	99,93

Le prix manquant en mars pour le produit 4 est maintenant imputé:

Prix	Déc-16	Jan-17	Fév-17	Mars-17	Avr-17
produit 1	6,00	6,30	6,20	6,40	6,00
produit 2	4,60	4,70	4,73	4,30	4,60
produit 3	5,20	5,40	5,20	5,30	5,20
produit 4	5,70	5,40	5,80	5,72	5,70

Un nouveau calcul de l'indice montre que l'indice d'avril est égal à celui de décembre. Par conséquent, l'indice satisfait au test d'identité.

Indices	Déc-16	Jan-17	Fév-17	Mars-7	Avr-17
produit 1		1,05	0,98	1,03	0,94
produit 2		1,02	1,01	0,91	1,07
produit 3		1,04	0,96	1,02	0,98
produit 4		0,95	1,07	0,99	1,00
Indice mensuel		1,01	1,01	0,99	1,00
Indice en chaîne	100,00	101,36	101,96	100,47	100,00

La comparaison directe entre les prix de mars et ceux de décembre a comme résultat un indice de 100,47. Ce chiffre est égal à l'indice en chaîne calculé ci-dessus: le test de transitivité est donc également satisfait.

3.3.2. Choix d'un indice de Jevons

Le calcul des indices élémentaires au niveau d'agrégation le plus bas s'effectue au moyen d'une moyenne géométrique ou indice de Jevons au niveau de chaque chaîne:

$$P_J = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{1/n} = \frac{\prod_{i=1}^{n} \left(p_i^t\right)^{1/n}}{\prod_{i=1}^{n} \left(p_i^0\right)^{1/n}} \qquad \textit{Équ. 2}$$

Nous avons choisi une moyenne géométrique (indice de Jevons) au lieu d'une moyenne arithmétique (indice de Dutot) pour la raison suivante. Un indice de Jevons présente l'avantage que le niveau des prix des observations n'a pas d'influence sur l'indice. Si un produit plus cher enregistre la même augmentation qu'un produit moins cher, ces deux hausses de prix auront le même poids dans le calcul de l'indice, ce n'est pas le cas avec l'indice de Dutot. Effectivement, ce dernier est sensible aux niveaux de prix observés. Le poids d'un prix dans le calcul d'une moyenne arithmétique est en effet proportionnel à son importance relative dans la moyenne arithmétique. En d'autres termes, l'indice de Dutot attribue plus de poids aux prix plus élevés qu'aux prix plus bas. C'est par exemple le cas des produits de marque plus coûteux par rapport aux produits meilleur marché. De même, dans le calcul de l'indice de Dutot, les plus grands conditionnements pèseront davantage dans la mesure de l'inflation que les plus petits conditionnements.

Vous trouverez un exemple ci-dessous:

Supposons que le produit B a un tout autre niveau de prix que les produits A et C. Et supposons que le produit A voit son prix diminuer de moitié, celui du produit B double et celui du produit C reste stable.

	Décembre	Janvier
Produit A	4	2
Produit B	20	40
Produit C	5	5
Moyenne arithmétique	9,67	15,67
Indice de Dutot	100	162,07

Compte tenu de l'évolution opposée des prix des produits A et B, l'inflation devrait rester stable. Cependant, dans une moyenne arithmétique, le produit B pèse plus que les produits A et C.

Si l'on utilise un indice de Jevons, on obtient le résultat suivant:

	Décembre	Janvier
Produit A	4	2
Produit B	20	40
Produit C	5	5
Moyenne géométrique	7,37	7,37
Indice de Jevons	100	100

L'indice de Jevons présente donc l'avantage de ne pas répercuter le niveau des relevés de prix sur l'évolution des prix mesurée. L'inflation enregistrée n'est dès lors pas biaisée par le choix de produits chers ou bon marché.

Il est également facile de démontrer que la moyenne géométrique de l'indice de Jevons est toujours inférieure à la moyenne arithmétique de l'indice de Dutot (sauf lorsque tous les prix sont égaux, les deux moyennes sont alors identiques). Prenons la moyenne arithmétique $a = \frac{x_1+x_2}{2}$, et la moyenne géométrique $b = \sqrt{x_1 x_2}$ de deux prix x_1 et x_2. Alors $b \leq a$ si $a - b \geq 0$. Si l'on complète les moyennes, on obtient:

$$a - b = \frac{x_1 + x_2}{2} - \sqrt{x_1 x_2} = \frac{1}{2}\left[\left(\sqrt{x_1}\right)^2 + \left(\sqrt{x_2}\right)^2 - 2\sqrt{x_1}\sqrt{x_2}\right] = \frac{1}{2}\left(\sqrt{x_1} - \sqrt{x_2}\right)^2$$

Ce résultat est toujours ≥ 0, ce qui prouve l'hypothèse.

L'indice de Jevons est également préférable d'un point de vue économique. Il suppose en effet une élasticité de substitution constante égale à 1. L'élasticité de substitution exprime la mesure dans laquelle les quantités varient proportionnellement par rapport aux variations de prix relatives. Dans l'indice de Jevons, l'élasticité de substitution est donc considérée comme constante. L'indice de Dutot suppose qu'il n'y a pas de substitution (élasticité de substitution de zéro), ce qui est illogique d'un point de vue économique.

En raison de ces caractéristiques statistiques et économiques, l'indice de Jevons a été privilégié.

3.3.3. Indice en chaîne et appariement des modèles

L'indice de Jevons obtenu au niveau élémentaire est donc un indice calculé avec des produits disponibles pendant deux périodes consécutives et pour lesquels la part des dépenses relatives au produit en question est supérieure au seuil dynamique déterminé. La formule de l'indice de Jevons est donc la suivante:

$$P_J = \prod_{i \epsilon G_{0,1}} \left(\frac{p_i^1}{p_i^0}\right)^{1/N_{0,1}} \qquad \text{Équ. 3}$$

où $G_{0,1}$ est l'échantillon, c'est-à-dire l'ensemble de tous les produits qui correspondent durant deux périodes consécutives (période 0 et période 1) et qui se trouvent au-dessus du seuil, et où $N_{0,1}$ est le nombre total de produits dans l'échantillon. Dans un indice en chaîne de plusieurs périodes, ces indices mensuels sont couplés en multipliant les indices obtenus. Pour un indice de Jevons en chaîne comprenant trois périodes, la formule est la suivante:

$$P_J^{02} = \prod_{i \epsilon G_{0,1}} \left(\frac{p_i^1}{p_i^0}\right)^{1/N_{0,1}} \prod_{i \epsilon G_{1,2}} \left(\frac{p_i^2}{p_i^1}\right)^{1/N_{1,2}} \qquad \text{Équ. 4}$$

Ces indices pour un groupe de produits k, exprimés dans l'année en cours avec comme période de base décembre de l'année précédente (=100) peuvent s'écrire $P_{J,k,dec=100}^{0t}$. Cela tient au fait que les coefficients de pondération sont toujours adaptés en décembre. Les indices obtenus sont ensuite agrégés avec les poids correspondants calculés à partir du chiffre d'affaires tiré des scanner data du groupe de produits concerné de l'année précédente:

$$P_A^{0t} = \frac{\sum_{k=1}^{n} w_k P_{J,k,dec=100}^{0t}}{\sum_{k=1}^{n} w_k} \qquad \text{Équ. 5}$$

L'adaptation annuelle des poids en décembre n'exerce aucun impact direct sur l'indice obtenu. Cependant, l'impact est indirect parce que cette adaptation a naturellement une influence sur la part de ce groupe de produits dans l'évolution des prix mesurée l'année suivante.

3.3.4. Pourquoi des indices non pondérés ?

Un indice non pondéré est utilisé au niveau élémentaire. Cela peut sembler étrange de prime abord, étant donné que les scanner data contiennent des informations relatives au chiffre d'affaires jusqu'au niveau des produits individuels. Grâce à ces informations, il serait possible de passer à une formule d'indice superlatif. Dans une formule d'indice superlatif, les prix et les quantités, tant de la période de référence que de la période considérée, sont traités de manière symétrique.

Pourquoi une formule d'indice superlatif n'est-elle pas utilisée et pourquoi les informations relatives au chiffre d'affaires ne sont-elles pas directement utilisées dans le calcul de l'indice, mais seulement de manière indirecte pour composer l'échantillon mensuel sur la base du seuil dynamique? Ceci est dû au fait que l'utilisation de ces informations relatives au chiffre d'affaires dans le calcul de l'indice entraîne un 'chain drift'. Un 'chain drift' signifie que l'indice en chaîne ne revient pas à l'unité (ou à un niveau d'indice de 100) quand les prix de la période considérée reviennent au niveau de la période de référence. Le concept du 'chain drift' sera expliqué à l'aide de l'indice de Törnqvist, P_T, dont la formule est égale à:

$$\begin{aligned} P_T^{0t} &= \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{0.5\left(\frac{p_i^0 q_i^0}{\sum_{j=1}^{n} p_j^0 q_j^0} + \frac{p_i^t q_i^t}{\sum_{j=1}^{n} p_j^t q_j^t}\right)} \\ &= \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{0.5\,(s_i^0 + s_i^t)} \end{aligned}$$

L'équation 6 exprime l'indice de Törnqvist pour la période de 0 à t et est égale à la moyenne géométrique des ratios de prix (période actuelle par rapport à la base), pondérée à partir de la moyenne arithmétique des parts de dépenses au cours de la période de base et de la période en cours. Pour un indice de Törnqvist en chaîne de deux périodes, la formule est la suivante:

$$P^{02}_{T,chainé} = \prod_{i=1}^{n} \left(\frac{p_i^1}{p_i^0}\right)^{0.5\,(s_i^0 + s_i^1)} \prod_{i=1}^{n} \left(\frac{p_i^2}{p_i^1}\right)^{0.5\,(s_i^1 + s_i^2)}$$

Supposons maintenant que seul 1 produit *a* est suivi (*i* = *a*). L'équation peut alors être ramenée à la forme suivante:

$$P^{02}_{T,chainé} = \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 + s_a^1)} \left(\frac{p_a^2}{p_a^1}\right)^{0.5\,(s_a^1 + s_a^2)}$$

Si ce produit *a* a fait l'objet d'une promotion temporaire lors de la période 1 ($p_a^0 > p_a^1$), mais revient à son niveau de prix de la période 0 lors de la période 2 ($p_a^0 = p_a^2$), alors l'équation 8 peut être réécrite comme suit:

$$\begin{aligned} P^{02}_{T,chainé} &= \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 + s_a^1)} \left(\frac{p_a^0}{p_a^1}\right)^{0.5\,(s_a^1 + s_a^2)} \\ &= \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 + s_a^1) - 0.5\,(s_a^1 + s_a^2)} \\ &= \left(\frac{p_a^1}{p_a^0}\right)^{0.5\,(s_a^0 - s_a^2)} \end{aligned}$$

Étant donné qu'en théorie (cf. test d'identité), l'indice doit revenir à l'unité parce que le niveau de prix de la période 2 est égal à celui de la période 0, cela ne peut se produire que lorsque $s_a^0 = s_a^2$, comme le montre l'équation 9. Cela signifie donc que les dépenses de la période 2 et de la période 0 doivent être égales vu que les niveaux de prix sont identiques pendant cette période. La théorie microéconomique classique part du principe que les quantités sont déterminées de manière unique pour un ensemble de prix donné. Si c'était le cas, les poids de la période 2 et de la période 0 seraient identiques.

Si cette prévision théorique ne se concrétise pas, alors, $P^{02}_{T,chainé} < 1$ ***lorsque*** $s_a^0 > s_a^2$ et $P^{02}_{T,chainé} > 1$ si $s_a^0 < s_a^2$. Un exemple fictif simple dans lequel l'indice de Törnqvist en chaîne ne revient pas à l'unité est présenté dans le tableau ci-dessous pour les produits A et B sur une période de quatre mois. Le prix et le chiffre d'affaires du produit B restent stables durant toute la période. Par contre, ceux du produit A fluctuent.

Tableau 8: Exemple fictif de 'chain drift' pour un indice de Törnqvist en chaîne

	Janvier		Février		Mars		Avril	
	Prix	Chiffre d'affaires	Prix	Chiffre d'affaires	Prix	Chiffre d'affaires	Prix	Chiffre d'affaires
Produit A	2,50	10.000	2,00	500.000	2,50	2.000	2,50	10.000
Produit B	3,00	10.000	3,00	10.000	3,00	10.000	3,00	10.000
Indice de Törnqvist en chaîne	100,00		84,78		96,35		96,35	

Dans cet exemple, le produit A est en promotion en février. Par conséquent, son chiffre d'affaires augmente fortement au cours de ce mois, entraînant ainsi un effet baissier sur l'indice (une diminution de 100 à 84,78). En mars, les prix ont retrouvé leur niveau de janvier, mais le chiffre d'affaires du produit A n'égale toutefois pas celui de janvier (par exemple, parce que les consommateurs ont constitué des stocks pendant la promotion de février). L'indice ne retrouve dès lors pas le niveau escompté de 100. De plus, il subsiste un effet permanent comme le montre l'indice d'avril. Afin de démontrer cet effet permanent, en avril, les données de chiffre d'affaires et de prix sont identiques à celles de janvier, ce qui, sans aucun doute, devrait déboucher sur un indice de 100. Ce n'est toutefois pas le cas. La valeur de l'indice reste de 96,35 étant donné que le rapport de prix ne change pas. Ce phénomène est appelé 'chain drift'.

Le graphique ci-dessous permet de visualiser le 'chain drift' de l'indice de Törnqvist en chaîne. Les chiffres de l'IPC mentionnés sont les indices officiels du produit X:

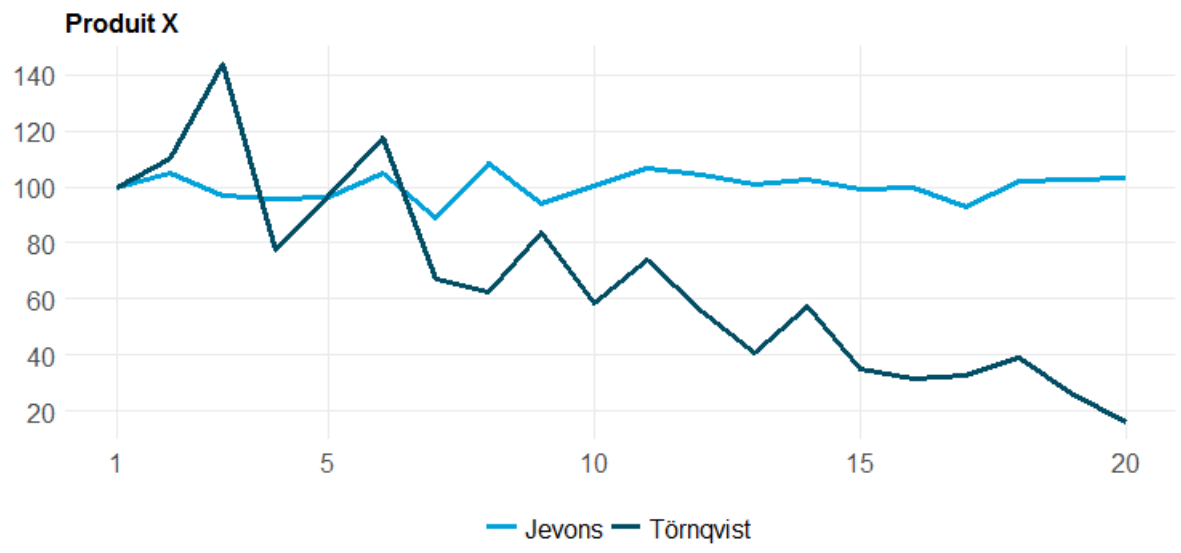


Figure 13: Chain drift d'un indice de Törnqvist en chaîne

Il apparaît clairement que l'indice de Törnqvist présente une dérive négative.

La question est désormais de savoir si ce 'chain drift' sera plutôt positif ou négatif. S'agit-il d'un indice qui affiche une tendance à la baisse (vers 0) sur de longues périodes ou d'un indice orienté à la hausse (vers l'infini)?

Dans un premier temps, on s'attend à ce que l'indice s'approche de zéro, parce que le composant faisant l'objet d'une baisse de prix reçoit un poids plus élevé en raison de la forte hausse de son chiffre d'affaires, et ce contrairement à la hausse des prix de la période suivante qui entraîne un poids moins important étant donné que les consommateurs achètent moins que ce à quoi on s'attendrait normalement (par exemple en raison de la constitution d'un stock ou d'une substitution). Cependant, il est également possible que la dérive soit positive, bien que cela se produise moins souvent en pratique. Une dérive positive peut se produire lorsqu'une promotion est d'application à la fin de la période de référence et se poursuit pendant une partie de la période considérée. La hausse du prix lors de la période considérée reçoit de cette manière un poids plus important, entraînant une dérive positive.

En raison du 'chain drift', il a donc été décidé d'utiliser un indice de Jevons non pondéré au niveau le plus bas. L'indice de Jevons est en réalité une variante particulière de l'indice de Törnqvist mentionné ci-dessus, dans laquelle les poids de tous les produits correspondent dans un agrégat élémentaire. Par conséquent, le poids d'un produit j pendant la période 0 $\left(s_j^0\right)$ est égal au poids durant la période t $\left(s_j^t\right)$. De même, les poids de tous les produits 1 à n sont égaux dans l'agrégat élémentaire. Cela signifie donc également que $s_j^0 = s_j^t = 1/n$.

Dans ce cas, la formule de Törnqvist est donc égale à:

$$P_T^{0t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{0.5\left(s_j^0 + s_j^t\right)} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{\left(s_j^0\right)} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{n}}, \qquad \textbf{Équ. 10}$$

ce qui donne l'indice de Jevons.

3.3.5. Filtres de dumping et des valeurs aberrantes

Un filtre de dumping et un filtre des valeurs aberrantes sont également appliqués avant le calcul de l'indice. Le filtre de dumping veille à ce que les produits qui enregistrent une forte baisse tant de leur prix que des quantités vendues puissent être exclus de l'échantillon. Cela permet d'éviter que ces produits influencent l'indice de manière négative (biais négatif). Cela tient au fait que, étant donné qu'il s'agit d'un indice en chaîne correspondant au modèle (voir ci-dessus), ces produits sortent de l'indice à un prix bas, mais n'y reviennent jamais. L'inclusion de ces produits dans le calcul engendrerait une dérive négative de l'indice.

De même, les produits qui connaissent chaque mois des variations de prix extrêmes sont exclus de l'échantillon par le filtre des valeurs aberrantes. Dans la pratique, le filtre des valeurs aberrantes retient principalement les produits qui peuvent être obtenus gratuitement au moyen d'une carte d'épargne.

L'exclusion à tort de produits par ces deux filtres n'a aucun impact à long terme sur l'indice, étant donné que les prix des produits exclus sont imputés. Il est donc préférable de faire preuve de trop de prudence en excluant des produits et en imputant leurs prix, plutôt que de travailler de manière imprudente et inclure dans le calcul des produits qui pourraient engendrer une dérive.

3.4. Rebranding et remplacements

Chaque mois, on établit également un lien entre les produits disparus et les nouveaux produits. Un produit peut en effet recevoir un autre code-barres, alors qu'il s'agit toujours du même produit. De même, un produit peut recevoir un autre code-barres car son contenu a changé (p.ex. X grammes supplémentaires) ou à la suite d'une promotion (p.ex. X% gratuit). Les produits reçoivent parfois un autre emballage, et donc un nouveau code interne. Celui-ci doit être lié à l'ancien code, étant donné qu'il s'agit toujours du même produit.

Deux exemples de produits qui ont été remplacés sont illustrés ci-dessous:

Mois	COICOP	Groupe	Ancien produit	Nouveau produit	Coeff.
1	01.1.1.4.02	Biscuits	Marque X - 6 + 2 gratuits - Chocolat	Marque X - 225 g - Chocolat	0.75
2	01.1.4.5.10	Fromage frais	Marque Y - 200 g - Fromage frais allégé	Marque Y - 300 g - Fromage frais allégé	1.50

Dans le premier exemple, l'ancien produit fait l'objet d'une promotion temporaire de deux biscuits gratuits. Le coefficient est alors de 0,75 (nouvelle quantité / ancienne quantité = 6 / 8). Dans le deuxième exemple, le contenu est modifié de 200 gr à 300 gr. Si le lien du premier exemple n'était pas établi, cela n'aurait pas un si grand impact sur l'indice à long terme, étant donné qu'il s'agit seulement d'une promotion temporaire. En revanche, le deuxième exemple peut, quant à lui, avoir un impact sur l'indice à long terme si ce lien n'est pas établi.

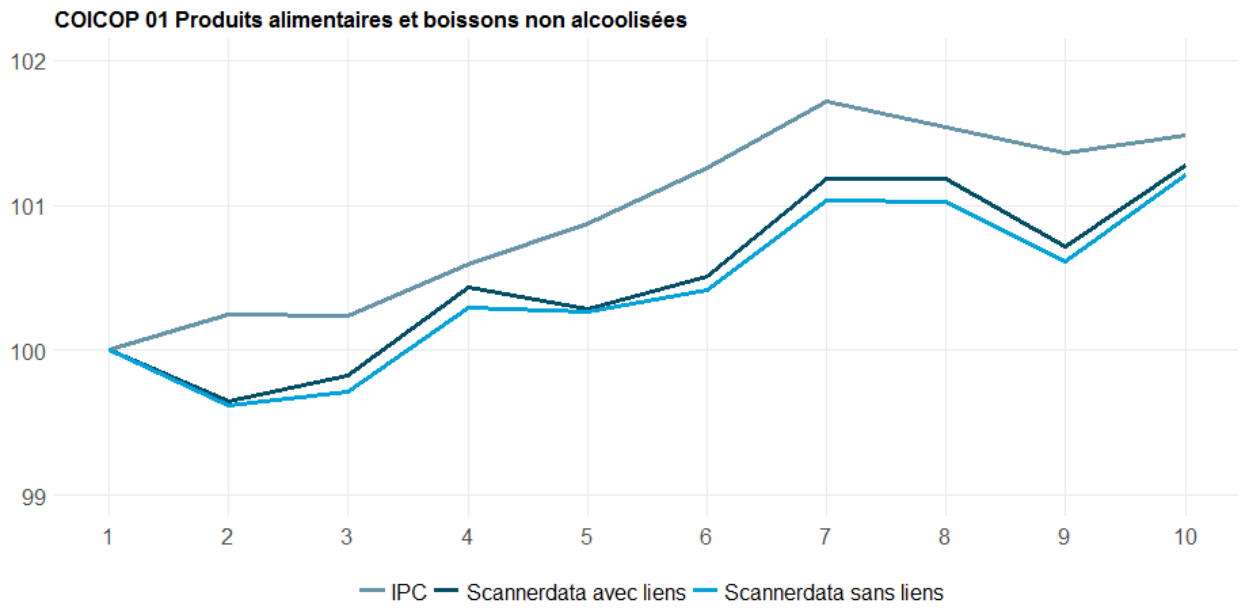


Figure 14: Influence du lien sur l'indice (COICOP 01)

Lors des relevés de prix classiques, l'enquêteur sait quel produit a été choisi le mois précédent et peut donc déterminer un successeur, avec un éventuel ajustement de la qualité ou de la quantité. Cette manière de procéder est appliquée aux scanner data.

Les remplacements et les relances de produits suivent généralement un certain schéma, comme le montre le graphique suivant.

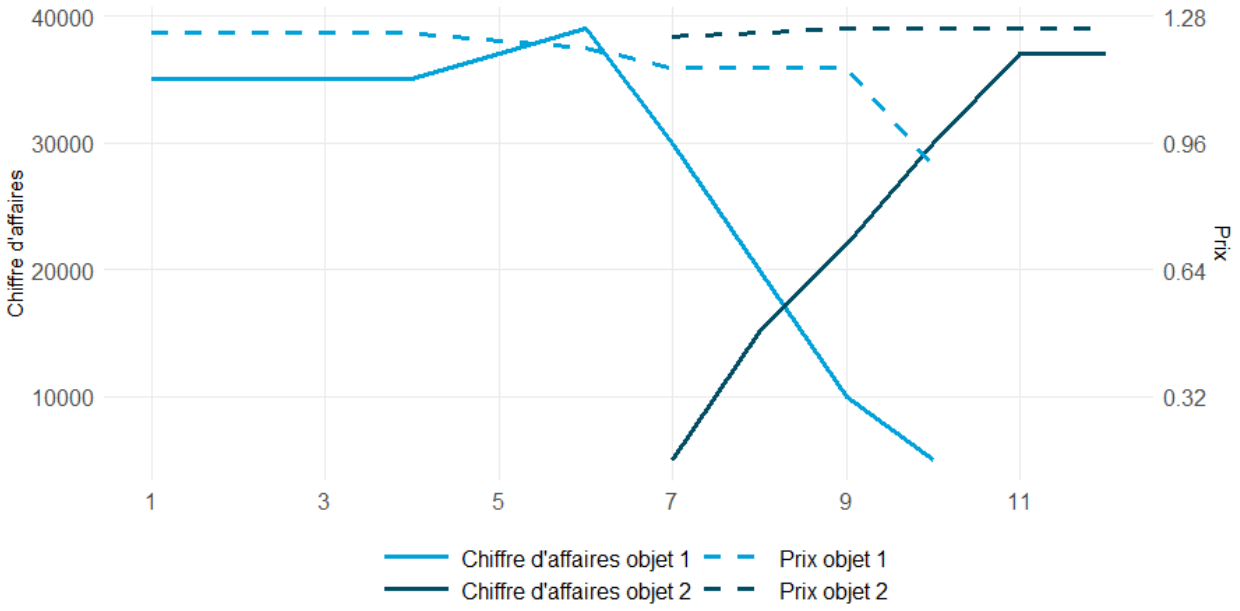


Figure 15: Exemple de remplacement et de relance de produits

Si un produit (objet 1) est remplacé par un nouveau produit (objet 2) un peu plus cher, la quantité vendue et le chiffre d'affaires diminuent pour l'ancien produit, et augmentent pour le nouveau produit.

La question est la suivante: pendant quel mois doit-on procéder au remplacement? Les produits ne restent représentatifs que si le chiffre d'affaires du nouveau mois est suffisamment important. Dans le cadre des relevés de prix traditionnels, l'enquêteur doit lui-même décider des produits qui sont représentatifs. Par conséquent, il est possible que des prix soient relevés tant que le produit est disponible, même si celui-ci n'est en réalité plus représentatif. Contrairement à la méthode traditionnelle, les scanner data fournissent davantage d'informations, ce qui permet de voir clairement quels produits sont représentatifs. La disponibilité des informations relatives au chiffre d'affaires met en évidence le fait que la représentativité de l'objet 2 augmente au détriment de l'objet 1. Il peut donc être décidé de procéder à un remplacement et de lier l'objet 1 (ancien produit) et l'objet 2 (nouveau produit). Les scanner data rendent ainsi les remplacements visibles.

Chaque mois, une liste est établie pour toutes les chaînes de supermarchés et pour tous les groupes de la COICOP 6. Cette liste reprend à la fois les «nouveaux» produits dans l'échantillon et les produits disparus. Pour les produits qui ont disparu, on examine le mois en cours et les trois mois précédents, étant donné que la disparition d'un produit et la mise sur le marché d'un nouveau (le successeur) ne surviennent pas nécessairement au même moment (le stock de l'ancien produit doit d'abord être vendu). Durant la période transitoire, la part de marché de l'ancien produit diminue et celui-ci disparaît de l'échantillon, alors que le nouveau produit affiche une part de marché croissante mais n'est pas encore repris dans cet échantillon.

Une application a été créée pour établir un lien entre les codes. De même, un coefficient est calculé de manière automatique à partir des scanner data pour les différences de quantité. Ce coefficient est égal au quotient de la nouvelle quantité et de l'ancienne. L'ancien prix est ensuite multiplié par ce coefficient afin de corriger cette différence de quantité. Si nécessaire, le coefficient peut être adapté. Si aucun lien n'est établi entre l'ancien et le nouveau produit, on peut supposer que la qualité a été modifiée et que le calcul se fait à l'aide d'un *bridged overlap* standard (analogue à l'imputation). Dans ce cas, l'évolution des prix de produits comparables sert de base pour estimer l'évolution du prix du nouveau produit et du produit remplacé. On estime donc le prix du nouveau produit (imputation) pour le mois précédent sur la base de l'évolution du prix de produits comparables entre le mois actuel et le mois précédent.

Outre la liste générale des nouveaux produits et des produits disparus, trois autres listes sont également créées. La première est la liste *perfect linking*. Elle contient les nouveaux produits et les produits disparus dont les descriptions sont identiques. Les produits de cette liste peuvent toujours être liés et ce lien est établi de manière automatique. La deuxième liste est la liste *reverse linking*. Cette liste montre les liens qui ont été établis par le passé et qui sont maintenant inversés. Prenons par exemple un produit faisant l'objet d'une promotion temporaire (6+2) au mois t et le produit redevenu normal (6) au mois t+1. Au mois t, le produit (faisant l'objet d'une promotion temporaire) est lié au produit (normal) au mois t-1. Le lien inverse (reverse) apparaît alors de manière automatique au mois t+1 sur cette liste et peut être encodé dans l'application. La troisième liste est la liste de *fuzzy linking*. Cette liste tente de lier des produits similaires sur la base de leurs descriptions (qui

ne sont pas identiques). Pour cela, les mots moins importants sont d'abord retirés de la description. Ensuite, un score est calculé sur la base du nombre de mots communs entre les deux produits (X et Y):

$$\text{score} = \frac{\#\ \text{mots commun}}{\sqrt{\#\ \text{mots X} * \#\ \text{mots Y}}}$$

Cette note se situe entre 0 et 1, et seules les scores supérieurs à 30 % sont reprises dans la liste de *fuzzy linking*.

3.5. Produits saisonniers

Les produits saisonniers sont des biens ou des services qui ne sont pas disponibles à l'achat – ou seulement en quantité limitée – pendant certains mois de l'année. Il doit s'agir d'un cycle type: les produits qui, par hasard, ne sont pas disponibles pendant un mois ne sont pas des produits saisonniers. Pour les scanner data des chaînes de supermarchés, il s'agit des produits frais (légumes frais, fruits frais et fruits de mer frais). D'autres produits saisonniers qui ne sont disponibles que pour une période très brève, comme les articles de Noël ou les œufs de Pâques, sont exclus du calcul de l'indice. En effet, ces produits ne sont disponibles que pour une très courte période. Par ailleurs, leur inclusion dans le calcul de l'indice engendrerait une dérive négative de celui-ci.

La dérive négative est occasionnée par l'utilisation dans le calcul de l'indice des prix relatifs de deux mois consécutifs (modèle correspondant, voir le chapitre 3.3.3). Pour ces articles, le prix utilisé dans le calcul de l'indice serait en réalité celui de la période suivant la «saison» et serait donc un prix de dumping (et non le prix représentatif pendant la «saison»). La probabilité que les mêmes produits reviennent l'année suivante est assez faible. Par conséquent, l'indice ne reviendrait pas au niveau initial.

Étant donné que tous les produits frais ne sont pas disponibles chaque mois, l'utilisation d'un panier dynamique est moins intéressante ici. Les groupes de la COICOP à 6 positions devraient alors être composés de manière très hétérogènes (par exemple un groupe mixte de fraises et bananes), car autrement, certains groupes ne contiendraient aucun produit pendant une période donnée (saison). Les groupes de la COICOP à 6 positions ont toutefois des poids fixes. Des imputations pendant plusieurs mois sont donc nécessaires, ce qui est contraire à la réglementation de l'IPCH. C'est pourquoi on utilise la méthode de pondération saisonnière au niveau de la classe au lieu d'un panier dynamique pour les produits frais. Cette méthode vise à limiter la variation au niveau des coefficients de pondération par mois et permet également d'attribuer un coefficient de pondération nul durant les mois où un produit n'est pas disponible sur le marché. Cela se traduit par les caractéristiques suivantes (un exemple fictif est présenté en annexe):

- un groupe de la COICOP à 6 positions a un poids nul lorsqu'il n'est pas disponible (hors saison);
- le poids moyen d'un groupe de la COICOP à 6 positions pendant l'année correspond à la part annuelle de ce groupe de produits dans le total des dépenses du groupe de produits du niveau supérieur. Cela veut donc dire que le poids moyen sur base annuelle des fraises dans le panier est égal à la part en pourcentage des fraises dans le total des dépenses des groupes de produits repris dans le groupe de la COICOP à 5 positions pour les fruits frais de toute l'année;
- le rapport des poids des groupes de la COICOP à 6 positions «en saison» reste égal tant que les deux groupes sont disponibles, également lorsque les poids individuels varient du fait que d'autres groupes sont ou non "en saison". Ainsi, le rapport de pondération des bananes et des fraises sera toujours le même lorsque les deux produits sont «en saison», même si d'autres produits "en saison" apparaissent ou si des produits "hors saison" disparaissent. La stabilité du ratio de pondération garantit que seule l'évolution pure des prix soit mesurée dans l'indice. Les modifications de poids de mois en mois n'ont aucun impact direct sur l'évolution des prix mesurée.

Les poids sont déterminés sur la base du chiffre d'affaires de l'année précédente. Pour déterminer si un groupe de la COICOP à 6 positions est "en saison" ou non, on examine les données des deux années précédentes. Cette méthode de pondération saisonnière au niveau de la classe est également utilisée pour les produits frais dans les relevés de prix classiques.

Les graphiques suivants montrent l'évolution de l'indice des groupes de la COICOP fruits frais et légumes frais, calculé à l'aide de la méthode classique et des scanner data. Pour les deux groupes de la COICOP, on remarque une évolution similaire entre les deux méthodes.

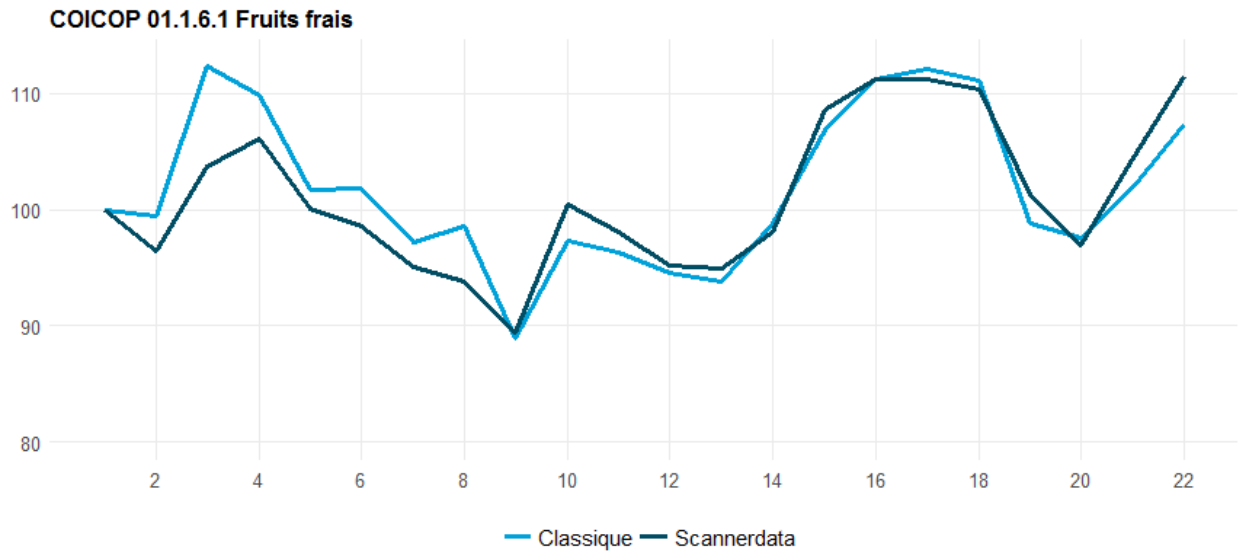


Figure 16: Comparaison de l'évolution des prix entre la méthode classique et les scanner data pour le groupe COICOP 01.1.6.1 Fruits frais

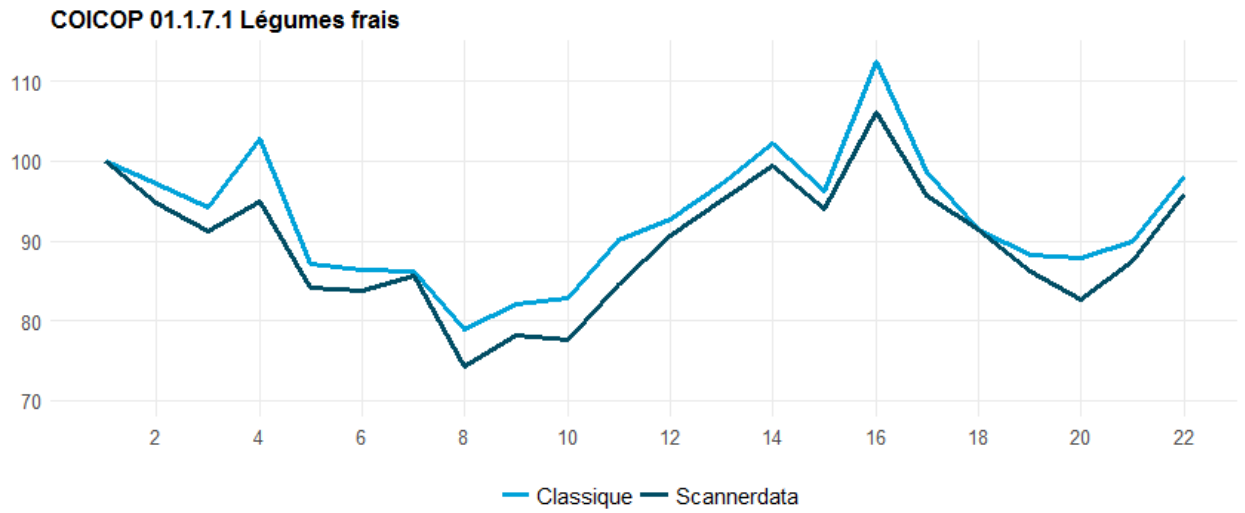


Figure 17: Comparaison de l'évolution des prix entre la méthode classique et les scanner data pour le groupe COICOP 01.1.7.1 Légumes frais

4. MODÈLE DE STRATIFICATION

Un modèle de stratification est utilisé pour combiner de manière effective les scanner data et les relevés de prix classiques. Les données sont combinées au niveau de la COICOP 5.

Une distinction est établie entre les "supermarchés et discounters" et les "magasins spécialisés" (bouchers, boulangers,...). Chacune des deux strates a un poids basé sur le comportement d'achat du consommateur. Ces poids sont mis à jour tous les deux ans à l'aide de l'enquête bisannuelle sur le budget des ménages.

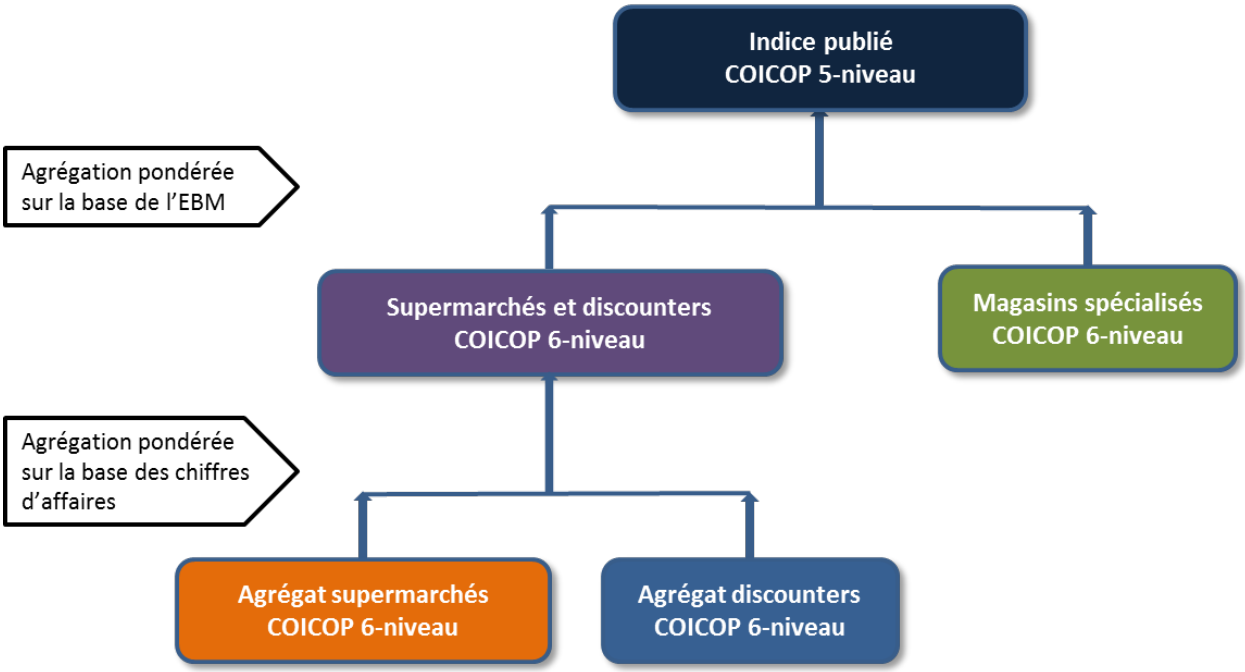


Figure 18: Modèle de stratification au niveau de la COICOP 5

La branche "supermarchés et discounters" est subdivisée en deux groupes, un pour les supermarchés (où les scanner data sont utilisées) et un pour les discounters (pour lesquels nous ne disposons pas de scanner data pour le moment). Chaque supermarché et discounter se voit attribuer un poids sur la base du chiffre d'affaires. Ces poids sont mis à jour chaque année.

Le groupe des supermarchés est subdivisé en trois chaînes dont nous recevons les scanner data. Chaque chaîne se voit attribuer un poids sur la base de ses comptes annuels (mise à jour annuelle). Comme déjà indiqué ci-dessus, des segments de consommation sont créés pour chaque chaîne. Chacun de ces groupes de la COICOP reçoit un poids sur la base des scanner data. Ce poids est renouvelé chaque année en fonction du chiffre d'affaires de l'année précédente. L'indice au niveau du segment de consommation est calculé à l'aide d'un indice de Jevons (moyenne géométrique).

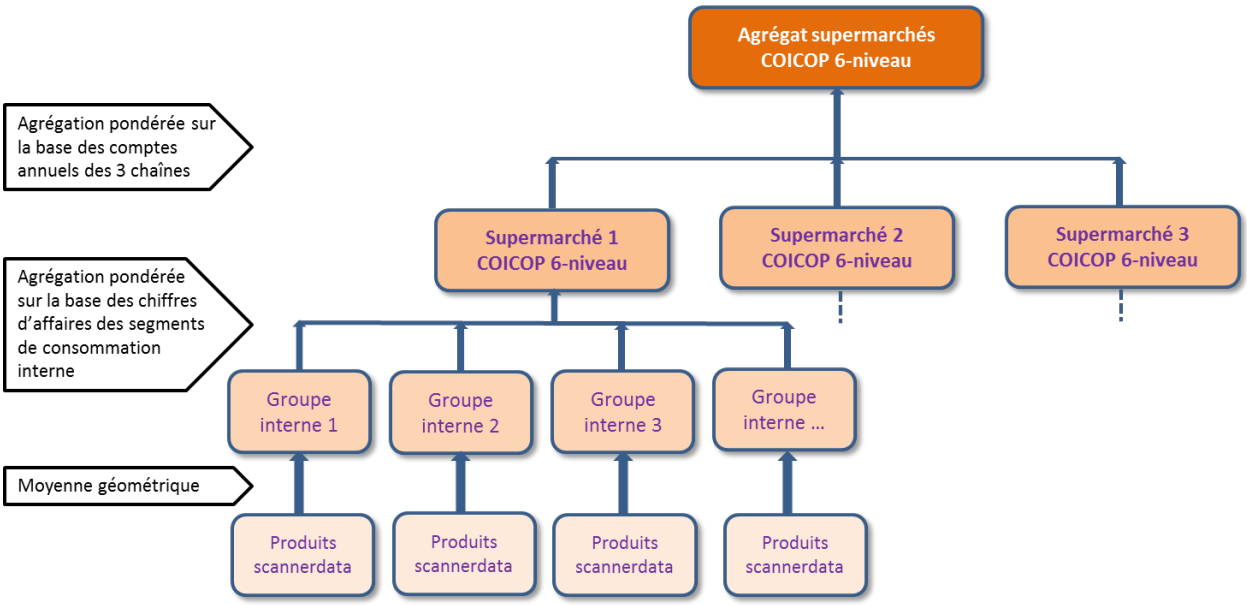


Figure 19: Modèle de stratification des chaînes de supermarchés au niveau de la COICOP 6

Il existe deux discounters (Aldi et Lidl) en Belgique. Leur poids est également mis à jour chaque année. Les prix sont collectés pour les agrégats traditionnels de la COICOP 6 à l'aide de la méthode classique. Ces agrégats se voient attribuer un poids sur la base de l'enquête sur le budget des ménages. Ces poids sont revus chaque année à la suite d'une mise à jour des prix ou de l'ajout d'un nouvel agrégat de la COICOP 6 . En ce qui concerne l'IPCH, le webscraping est utilisé pour un discounter.

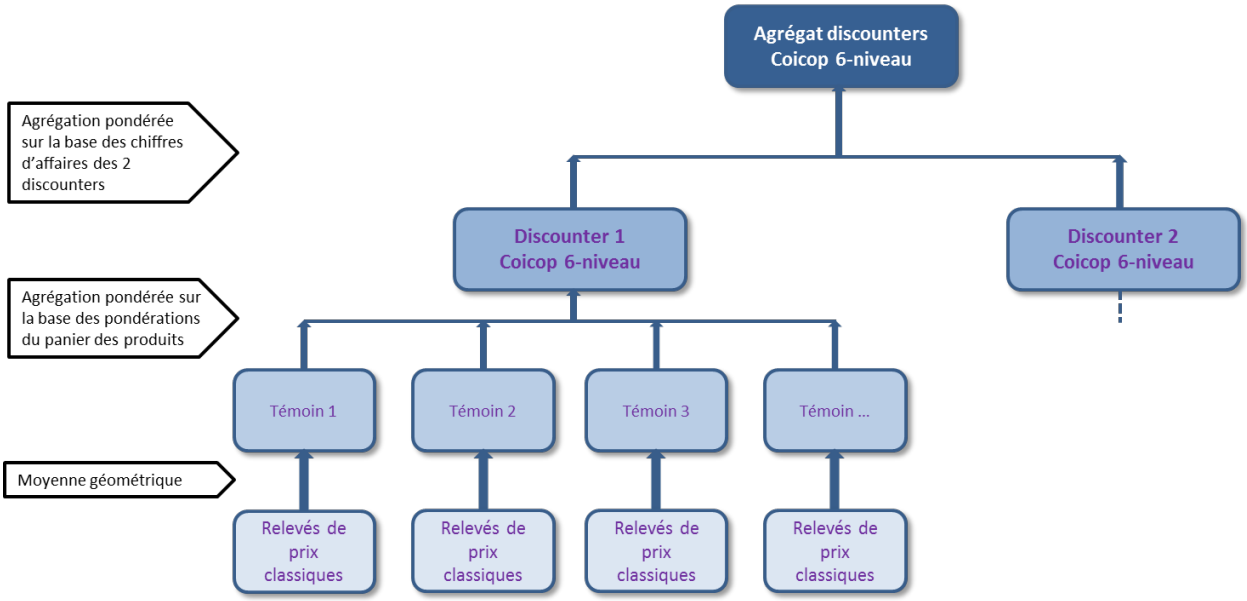


Figure 20: Modèle de stratification des discounters au niveau de la COICOP 6

Outre les supermarchés et les discounters, il existe également des magasins spécialisés, où les enquêteurs relèvent les prix selon la méthode traditionnelle. Un indice de Jevons est alors utilisé pour l'agrégation. Les poids sont ensuite appliqués au niveau de la COICOP 6, sur la base de l'EBM.

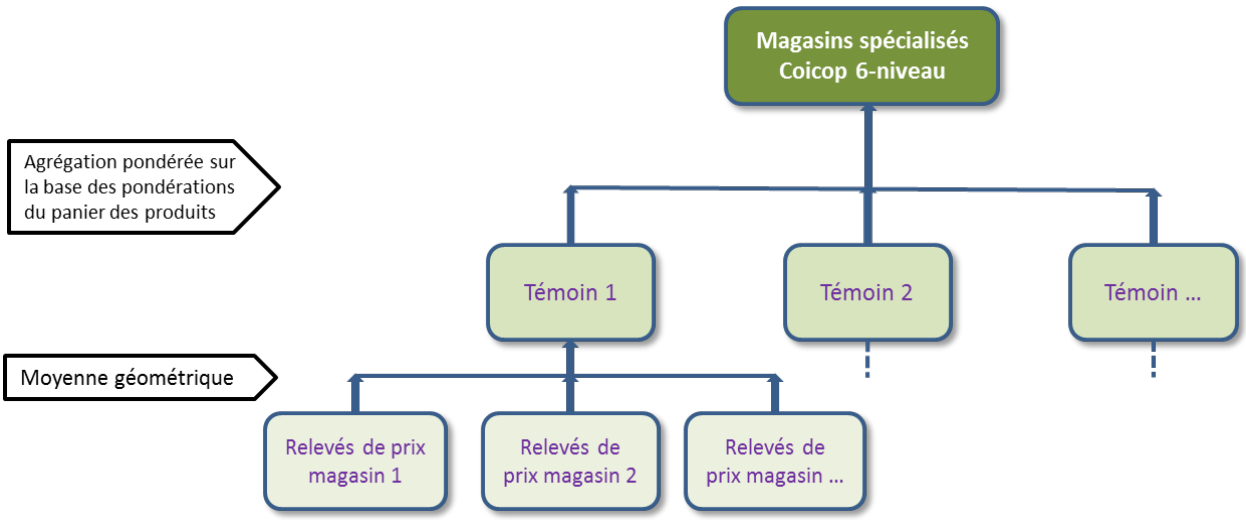


Figure 21: Stratification des magasins spécialisés au niveau de la COICOP 6

Cependant, pour les magasins spécialisés, les prix ne sont relevés que lorsque la part de marché est (selon l'EBM) supérieure à 10 %. Le tableau ci-dessous donne un aperçu des groupes de la COICOP pour lesquels les prix sont encore relevés de manière traditionnelle.

COICOP	Groupe	Supermarchés et discounters	Magasins spécialisés
01.1.1.3	Pain	53,80%	46,20%
01.1.2.8	Autres préparations à base de viande	74,37%	25,63%
01.1.2.1	Viande de bœuf et de veau	76,23%	23,77%
01.1.2.5	Autres viandes	76,45%	23,55%
01.1.2.2	Viande porcine	77,67%	22,33%
01.1.2.7	Charcuteries	78,55%	21,45%
01.1.1.4	Autres produits de boulangerie	80,19%	19,81%
01.1.2.3	Viande de mouton et d'agneau	80,87%	19,13%
01.1.2.4	Volaille	84,35%	15,65%
01.1.3.1	Poisson frais	87,73%	12,27%

Pour les groupes dont la part de marché des magasins spécialisés est inférieure à 10 % (par exemple les légumes et les fruits), on observe une forte corrélation entre l'indice basé sur les prix des supermarchés et l'indice basé sur les prix des magasins spécialisés. Un exemple est présenté ci-dessous pour le groupe COICOP 01.1.7.1 Légumes frais hormis pommes de terre. On peut en conclure que l'évolution des prix mesurée entre les chaînes de supermarchés et les magasins spécialisés est presque identique.

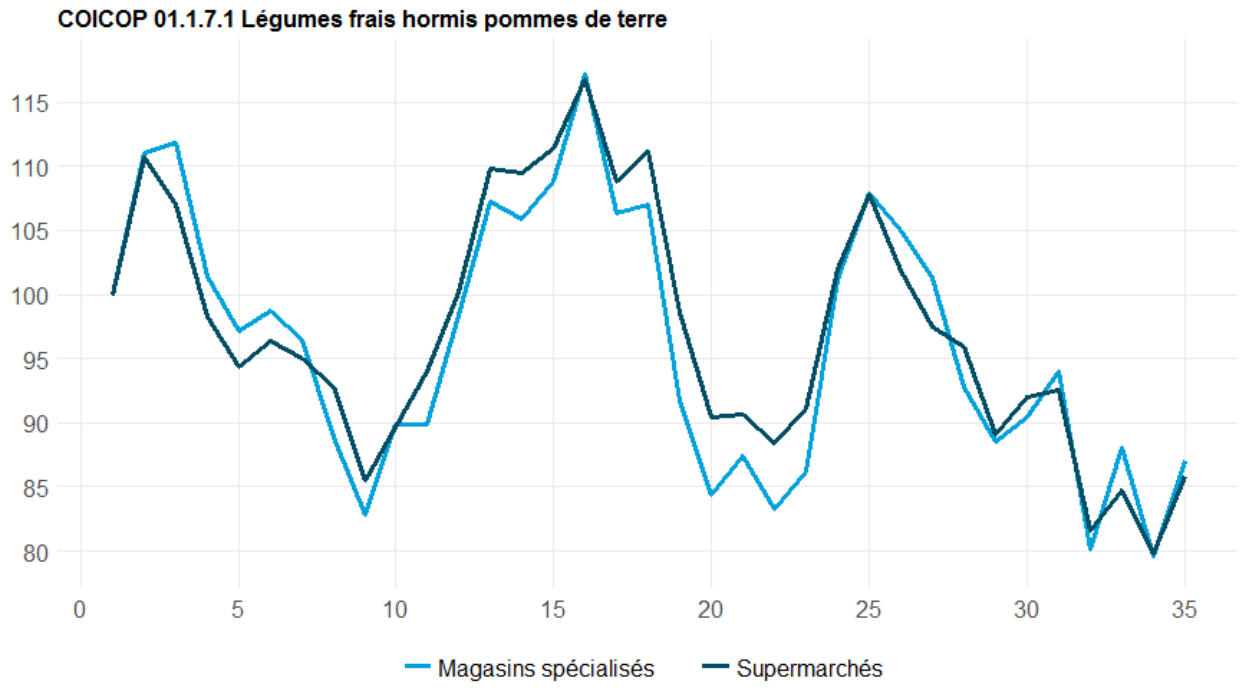


Figure 22: Comparaison de l'évolution des prix entre les chaînes de supermarchés et les magasins spécialisés

5. CONCLUSION

Statbel reçoit les données de trois chaînes de supermarchés via SFTP. Une liste contenant les nouveaux produits, qui doivent être liés au bon groupe COICOP, est automatiquement créée. Un segment de consommation COICOP est proposé sur la base de la classification interne. L'apprentissage automatique permet de créer une deuxième liste contenant des suggestions de groupes COICOP. Après vérification, un fichier contenant les codes produit et leur classification COICOP est créé. Ensuite, un indice préliminaire est calculé (à partir des scanner data) afin de déterminer le panier dynamique. Les listes permettant de lier les relances sont également créées. Les relances de produits correctes sont ensuite chargées dans le datawarehouse. L'indice définitif peut alors être calculé sur la base d'un indice de Jevons en chaîne. Les indices obtenus sur la base des scanner data sont intégrés dans le calcul complet de l'indice au moyen d'un modèle de stratification.

L'utilisation des scanner data a amélioré la mesure de l'inflation. L'utilisation des informations relatives au chiffre d'affaires, disponibles dans les scanner data, permet d'inclure presque tous les produits représentatifs dans le calcul de l'indice via le panier dynamique. Les poids peuvent également être déterminés de manière plus précise au niveau le plus bas dans le calcul de l'indice. Un simple exemple permet d'illustrer cela. Dans le groupe du « sucre », un seul produit faisait l'objet d'un suivi jusqu'à la fin 2015 (certes pour différentes marques et différents points de vente). Il s'agissait du sucre cristallisé en emballage d'un kilo. Depuis janvier 2016, un suivi de l'évolution des prix est réalisé pour tous les types de sucre et tous les conditionnements qui obtiennent un chiffre de vente représentatif. Autrement dit, outre l'évolution du prix du sucre cristallisé, un suivi est également réalisé et intégré au sous-indice du groupe du « sucre » pour les témoins suivants : sucre en morceaux, sucre de canne, sucre candi, cassonade, sucre glace, sucre liquide, etc. De plus, ce suivi de l'évolution des prix s'applique aux différents conditionnements (250 g, 500 g, 750 g, 1 kg…). Il est évident que cette méthode contribue fortement à améliorer la représentativité de l'indice des groupes concernés.

En ce qui concerne les autres chaînes de supermarchés et les petits détaillants pour lesquels les scanner data ne sont pas encore transmises, les relevés de prix classiques en magasin (ou le webscraping) restent en partie nécessaires pour obtenir une image fidèle de l'évolution de prix totale en Belgique. Les scanner data sont aussi actuellement utilisées pour maintenir la représentativité de l'échantillon de produits de ces relevés classiques, et ont donc ainsi également un impact. Des discussions sont également en cours avec les autres chaînes de supermarchés afin de pouvoir aussi traiter leurs scanner data. Enfin, Statbel étudie également un élargissement à d'autres secteurs.

La méthode actuelle n'utilise pas encore les informations de pondération explicites au niveau le plus bas, mais ce point est examiné par Statbel et d'autres offices de statistique afin de pouvoir tout de même éventuellement les utiliser au niveau le plus bas tout en évitant un "chain drift". Ces méthodes ne sont pas prises en compte dans la présente analyse, étant donné qu'elles sont toujours en cours d'examen.

ANNEXE

Le tableau ci-dessous donne un aperçu de tous les groupes COICOP pour lesquels des scanner data sont utilisées ainsi que leur poids correspondant (2017):

COICOP	Dénomination	Pondérations publiées 2017 (‰)
01.1.1.1	Riz	0,43
01.1.1.2	Farines et autres céréales	0,77
01.1.1.3	Pain	9,94
01.1.1.4	Autres produits de boulangerie	10,64
01.1.1.5	Pizza et quiche	1,53
01.1.1.6	Pâtes alimentaires et couscous	2,55
01.1.1.7	Céréales du petit déjeuner	1,29
01.1.1.8	Autres produits à base de céréales	0,61
01.1.2.1	Viande de bœuf et de veau	6,02
01.1.2.2	Viande de porc	2,62
01.1.2.3	Viande de mouton et d'agneau	1,03
01.1.2.4	Volaille	4,81
01.1.2.5	Autres viandes	0,95
01.1.2.7	Charcuteries (viande salée, séchée ou fumée)	9,57
01.1.2.8	Autres préparations à base de viande	14,08
01.1.3.1	Poisson frais	3,54
01.1.3.2	Poisson surgelé	0,82
01.1.3.3	Fruits de mer frais	2,01
01.1.3.4	Fruits de mer surgelés	0,66
01.1.3.5	Poisson et fruits de mer fumés	1,17
01.1.3.6	Poisson en conserve et préparations à base de poisson et de fruits de mer	2,37
01.1.4.1	Lait entier	0,88
01.1.4.2	Lait écrémé et demi-écrémé	1,62
01.1.4.3	Lait concentré ou en poudre	0,17
01.1.4.4	Yaourt	2,48
01.1.4.5	Fromage	9,76
01.1.4.6	Autres produits laitiers	2,52
01.1.4.7	Oeufs	1,11
01.1.5.1	Beurre	2,09
01.1.5.2	Margarine et graisses végétales	1,14
01.1.5.3	Huile d'olive	0,75
01.1.5.4	Autres huiles alimentaires	0,59
01.1.6.1	Fruits frais	10,69
01.1.6.3	Fruits séchés et à coque	1,13
01.1.6.4	Fruits en conserves	0,38
01.1.7.1	Légumes frais hormis pommes de terre	9,57
01.1.7.2	Légumes surgelés hormis pommes de terre	0,61
01.1.7.3	Légumes séchés, en conserve et préparations à base de légumes	2,26
01.1.7.4	Pommes de terre	3,23
01.1.7.5	Chips	1,17
01.1.8.1	Sucre	0,46
01.1.8.2	Confiture, marmelade et miel	1,18
01.1.8.3	Chocolat	4,72

01.1.8.4	Confiseries	2,10
01.1.8.5	Glace et sorbet	1,52
01.1.9.1	Sauces et condiments	2,20
01.1.9.2	Sel, épices et plantes aromatiques	1,03
01.1.9.3	Aliments pour nourrisson	1,06
01.1.9.4	Plats préparés	1,40
01.1.9.9	Autres produits alimentaires n.c.a. (y compris préparations diététiques)	1,57
01.2.1.1	Café	3,26
01.2.1.2	Thé	0,63
01.2.1.3	Cacao et poudre à base de chocolat	0,14
01.2.2.1	Eau minérale ou de source	3,47
01.2.2.2	Boissons rafraîchissantes	7,35
01.2.2.3	Jus de fruits et de légumes	2,40
02.1.1.1	Liqueurs et eaux-de-vie	2,08
02.1.1.2	Boissons rafraîchissantes à base d’alcool	0,07
02.1.2.1	Vin	8,61
02.1.2.2	Vin d'autres fruits	0,11
02.1.2.3	Vin liquoreux	0,91
02.1.3.1	Bière légère	2,40
02.1.3.2	Bière forte	2,01
02.2.0.1	Cigarettes	6,92
02.2.0.3	Autres produits du tabac	2,04
05.5.2.2	Accessoires divers pour la maison et le jardin	3,03
05.6.1.1	Produits de nettoyage et d'entretien	6,45
05.6.1.2	Autres petits articles de ménage non durables	4,32
09.3.4.2	Produits pour animaux de compagnie	6,70
09.5.4.1	Produits de papier	1,05
09.5.4.9	Matériel pour écrire et dessiner	1,58
12.1.3.1	Appareils non électriques pour soins corporels	0,93
12.1.3.2	Articles d'hygiène corporelle et produits de beauté	16,11

Le tableau ci-dessous présente un exemple fictif des poids pour les fruits frais:

	Jan	Fév	Mars	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc	Moy.
Fruits frais	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Pommes: Jonagold	3%	3%	3%	2%	2%	2%	1%	2%	1%	2%	2%	2%	2%
Pommes: Golden	4%	4%	4%	3%	3%	2%	2%	2%	2%	3%	3%	3%	3%
Pommes: Granny	16%	16%	16%	12%	12%	8%	7%	9%	7%	12%	12%	12%	12%
Poires rondes	4%	4%	4%	3%	3%	2%	2%	2%	2%	3%	3%	3%	3%
Poires: Conférence	3%	3%	3%	3%	3%	0%	0%	0%	1%	2%	2%	2%	2%
Raisins	0%	0%	0%	0%	0%	0%	0%	0%	16%	26%	26%	26%	8%
Pêches	0%	0%	0%	0%	0%	6%	5%	7%	6%	0%	0%	0%	2%
Abricots	0%	0%	0%	0%	0%	6%	5%	0%	0%	0%	0%	0%	1%
Prunes	0%	0%	0%	0%	0%	0%	0%	13%	11%	0%	0%	0%	2%
Cerises	0%	0%	0%	0%	0%	13%	12%	0%	0%	0%	0%	0%	2%
Nectarines	0%	0%	0%	0%	0%	0%	14%	20%	16%	0%	0%	0%	4%
Oranges	19%	19%	19%	15%	15%	9%	8%	11%	9%	14%	14%	14%	14%
Citrons	3%	3%	3%	2%	2%	1%	1%	2%	1%	2%	2%	2%	2%
Pamplemousses	3%	3%	3%	2%	2%	1%	1%	1%	1%	2%	2%	2%	2%
Mandarines	17%	17%	17%	0%	0%	0%	0%	0%	0%	13%	13%	13%	8%
Bananes	18%	18%	18%	14%	14%	9%	7%	10%	8%	14%	14%	14%	13%
Fraises	0%	0%	0%	37%	37%	23%	19%	0%	0%	0%	0%	0%	10%
Melons	0%	0%	0%	0%	0%	14%	12%	16%	13%	0%	0%	0%	5%
Kiwis	9%	9%	9%	7%	7%	4%	4%	5%	4%	6%	6%	6%	6%

À PROPOS DE STATBEL

Statbel, l'office belge de statistique, collecte, produit et diffuse des chiffres fiables et pertinents sur l'économie, la société et le territoire belges.

La collecte s'effectue à l'aide de sources de données administratives et d'enquêtes. La production est réalisée de manière qualitative et scientifique. Les statistiques sont diffusées en temps opportun et de manière conviviale.

Statbel garantit que, d'une part, la vie privée et les données confidentielles sont protégées et que, d'autre part, les données sont utilisées à des fins exclusivement statistiques.

Visitez notre site internet
www.statbel.fgov.be

ou contactez-nous
e-mail: statbel@economie.fgov.be

Statbel (Direction générale Statistique - Statistics Belgium)
North Gate - Boulevard du Roi Albert II, 16, 1000 Bruxelles
E-mail: statbel@economie.fgov.be

Numéro d'entreprise
0314.595.348

Editeur responsable
Nicolas Waeyaert

North Gate
Boulevard du Roi Albert II, 16
1000 Bruxelles