STATBEL
La Belgique en chiffres

Le webscraping, la collecte et le traitement de données en ligne pour l'indice des prix à la consommation

- Ken Van Loon et Dorien Roels -

n°02

ANALYSE

01.2018

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

.be

Le webscraping, la collecte et le traitement de données en ligne pour 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 du webscraping dans l'indice des prix à la consommation. En quoi consiste le webscraping? À quoi ressemblent ces données? Et comment Statbel (DG Statistique – Statistics Belgium) traite-t-elle ces données?

Vu l'importance croissante des boutiques en ligne et l'émergence des ventes en ligne des 'magasins classiques', Statbel estime qu'il est nécessaire d'inclure ces données dans le calcul de l'indice des prix à la consommation. Par ailleurs, l'avantage du webscraping par rapport aux relevés de prix classiques est qu'il permet de suivre le prix d'un nombre beaucoup plus important de produits.

Le webscraping est une technique qui permet d'extraire automatiquement des données de pages internet (' to scrap'). Statbel utilise le langage de programmation R pour effectuer le webscraping. Les scripts programmés sont exécutés de manière automatique à intervalles réguliers, et les données collectées sont sauvegardées de manière centralisée pour un traitement ultérieur. Une application de tableau de bord a été conçue pour contrôler l'exécution des scripts. Un outil robotisé a également été créé pour suivre les variations de prix de certains produits spécifiques.

Ces derniers mois, Statbel a réalisé toutes sortes d'études de cas. En effet, pour certains produits comme les produits électroniques grand public, les chaussures, les hôtels, les voitures d'occasion et les chambres d'étudiant, les données peuvent être collectées en ligne. Statbel a récolté les données des groupes de produits visés ci-dessus et a testé différentes méthodes de calcul de l'indice sur la base du webscraping.

Il ressort des résultats des études de cas sur les produits électroniques grand public et les voitures d'occasion que les caractéristiques qualitatives permettent d'effectuer des calculs difficiles à réaliser avec la méthode de calcul traditionnelle. En ce qui concerne le groupe de produits des hôtels, le webscraping permet par exemple d'élargir l'échantillon, ce qui favorise la qualité des données. Et l'étude de cas sur les chaussures démontre que le webscraping donne des résultats comparables à ceux des relevés de prix classiques effectués dans des magasins physiques.

Le présent article décrit également un certain nombre d'algorithmes pour l'apprentissage automatique. L'apprentissage automatique est l'étude dans le cadre de laquelle des algorithmes sont créés pour que les machines/ordinateurs/programmes puissent "apprendre" eux-mêmes. Cela signifie que non seulement des actions préprogrammées sont effectuées, mais également que le programme peut se développer lui-même. Par exemple, grâce à des modèles obtenus à partir des données classifiées - l'apprentissage automatique supervisé - le programme apprend à classer de nouvelles données. Cette classification s'effectue sur la base de la reconnaissance de mots et de codes. Différents algorithmes ont été testés et Statbel utilise déjà actuellement l'apprentissage automatique supervisé dans le cadre des scanner data.

SOMMAIRE

1. INTRODUCTION

L'indice des prix à la consommation (IPC) est une statistique mensuelle établie par Statbel (la DG Statistique – Statistics Belgium du SPF Economie). 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 la technique du *webscraping* et du traitement des données ainsi collectées. Le *webscraping* est une technique qui permet d'extraire automatiquement des données de pages internet *(' to scrap')*. Cependant, les pages Web sont composées de codes HTML et contiennent des quantités gigantesques de données textuelles. Il faut donc collecter et traiter ces données de manière structurée afin qu'elles puissent être utilisées à des fins statistiques. La mise en œuvre du *webscraping* nécessite donc un autre type de compétences, en particulier la programmation, la collecte et le traitement des données. Dans le contexte de l'indice des prix à la consommation, il est intéressant de collecter les prix et les descriptions des produits sur les boutiques en ligne.

Vu l'importance croissante des boutiques en ligne, ainsi que des ventes en ligne des "magasins classiques", il est nécessaire d'aussi inclure ces données dans le calcul des indices des prix. L'utilisation de ces données permet également d'améliorer l'efficacité de la collecte des données. Elle accroit également la qualité des statistiques.
Cet article décrit comment y parvenir. Il est structuré comme suit:

- introduction aux concepts "big data" et webscraping;
- comment Statbel exécute-t-elle le webscraping?
- évolution des prix des produits électroniques grand public et différentes méthodes de calcul de l'indice pour ce type de données;
- études de cas et méthodes: voitures d'occasion, chambres d'hôtel, chambres d'étudiant et chaussures;
- apprentissage automatique et méthodes de classement automatique des données issues du webscraping ou d'exclusion de certaines observations;
- tableau de bord du webscraping permettant de contrôler les scripts et outil robotisé permettant d'observer les variations de prix de certains produits;
- état des lieux.

Pour d'autres statistiques également, le webscraping offre une alternative intéressante à l'interrogation directe classique. Dans ce cadre, Statbel étudie notamment la possibilité de mesurer le nombre de postes vacants à l'aide de cette technique. L'utilisation du webscraping à d'autres fins que l'IPC ne relève toutefois pas du champ d'application de la présente analyse.

2. BIG DATA

On peut se poser la question de savoir si, à vrai dire, le *webscraping* fait partie des *big data*. Les datasets obtenus ne contiennent en effet pas des milliards d'entrées, mais sous l'angle du concept de statistique des prix, le *webscraping* peut être considéré comme une forme de *big data*. Par exemple, avec les relevés de prix classiques, on obtient environ 15.000 prix par an pour les chaussures, alors que le *scraping* quotidien du site internet d'un seul magasin permet déjà d'obtenir plus d'un million de prix par an. Le traitement de ces données pour obtenir des indices doit dès lors s'effectuer de manière plus automatisée qu'avec les relevés de prix classiques. On ne peut donc pas appliquer à ces données en ligne les méthodes utilisées pour le traitement des données des relevés de prix classiques.

Bien entendu, avoir plus de données ne signifie pas pour autant que les données sont meilleures. L'objectif ultime doit donc être de déterminer si les prix en ligne peuvent être utilisés pour estimer l'inflation, éventuellement en tant que proxy des prix hors ligne ou en complément des prix traditionnels.

Un avantage potentiel est cependant que la fréquence de la collecte des prix peut logiquement être augmentée grâce au *webscraping*. Par exemple, pour les vêtements et les chaussures, les prix ne sont actuellement pas collectés partout chaque mois. La fréquence quotidienne du *webscraping* représente évidemment une amélioration. Les prix en ligne sont déjà utilisés actuellement comme proxy des prix hors ligne via la collecte manuelle de données. Par exemple, la collecte des prix de nombreux produits électroniques grand public est aussi déjà effectuée en ligne sans *webscraping*. Les prix sont dans ce cas notés de manière centralisée à partir des sites internet des commerçants. Cette méthode est utilisée non seulement en Belgique mais aussi dans la plupart des pays européens.

Le Billion Prices Project[2] du MIT (Massachusetts Institute of Technology) a également montré que l'inflation estimée (quotidiennement) sur la base des prix en ligne était très proche de l'inflation officielle calculée par le Bureau of Labor Statistics.

Actuellement, ce projet est commercialisé par une spin-off (PriceStats[3]) et l'inflation et les parités de pouvoir d'achat est estimée pour 22 pays (dont nos pays voisins). Il convient toutefois de noter que tant le Billion Prices Project que PriceStats utilisent les chiffres officiels de l'inflation pour les segments où il est difficile de mesurer les prix en ligne (essentiellement les services et l'énergie). Ils ne communiquent pas où les chiffres officiels et où seuls les prix en ligne sont utilisés.

Il convient également de noter que le *webscraping* ne pose pas problème juridique en ce qui concerne l'indice des prix à la consommation. En effet, la réglementation relative à l'indice des prix à la consommation harmonisé européen (IPCH)[4] stipule que les unités statistiques fournissent, si nécessaire, des données pour le calcul de l'IPCH :

> *3. Les unités statistiques qui communiquent des informations sur les produits inclus dans les dépenses monétaires de consommation finale des ménages coopèrent à la collecte ou à la communication des informations de base selon les besoins. Les unités statistiques sont tenues de transmettre des informations de base exactes et complètes aux organismes nationaux chargés du calcul des indices harmonisés.*[5]

[2] http://www.thebillionpricesproject.com/
[3] https://www.pricestats.com/
4 Outre l'indice national des prix à la consommation (IPC), Statbel calcule également l'indice européen des prix à la consommation harmonisé (Harmonised Index of Consumer Prices, HICP). L'IPCH permet de comparer les taux d'inflation des États membres de l'Union européenne. L'optique des dépenses et les méthodes appliquées sont coordonnées autant que possible et définies dans la réglementation européenne. Les résultats de l'IPC et de l'IPCH ne sont 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.
5 Règlement (UE) 2016/792 article 5.3

3. IMPLICATIONS POUR LE CALCUL DE L'IPC

Outre le fait que le *webscraping* est assez technique (voir chapitre 4) et nécessite un certain nombre de compétences en programmation, l'approche du calcul de l'indice est également différente.

La collecte de données classique est principalement basée sur un échantillon ciblé (échantillonnage dirigé). Un certain nombre d'éléments représentatifs ("témoins") sont préalablement sélectionnés par segment de consommation. Une définition est ensuite établie pour ces éléments. Les prix sont enregistrés par les enquêteurs sur la base de cette définition. Selon le produit, ces prix sont enregistrés localement dans les magasins ou de manière centralisée. Il s'agit donc d'une approche de bas en haut (*bottom-up*), dans laquelle on définit d'abord de manière centralisée les produits dont les prix sont nécessaires, alors qu'avec le *webscraping*, aucune sélection préalable n'est effectuée. L'approche est davantage de haut en bas (*top-down*) car on part du site internet d'un détaillant, d'un site de réservation ou d'un site de petites annonces et la quasi-totalité des prix sont collectés, après quoi ils sont analysés et classifiés afin de pouvoir effectuer les calculs nécessaires. Cela peut éventuellement s'effectuer par apprentissage automatique (*machine learning*) (voir explications plus détaillées au chapitre 10).

Étant donné que l'indice des prix à la consommation est calculé sur une base mensuelle, les changements apportés à un site internet doivent être détectés rapidement afin que le script qui réalise le scraping d'un site donné puisse être rapidement adapté. Afin de pouvoir détecter rapidement ces changements, une application de tableau de bord a été développée. Elle est décrite plus en détail au chapitre 11.

4. COMMENT FONCTIONNE LE *WEBSCRAPING*?

Chaque site internet est conçu différemment. Différents scripts sont dès lors programmés pour extraire les données des différents sites. Dans un premier temps, il faut observer la structure du site. Il faut ensuite regarder les différents groupes de produits et les éventuelles sous-catégories. Une fois qu'elles ont été déterminées, ces catégories peuvent être parcourues pour effectuer le *scraping* de tous les prix et caractéristiques de produit correspondants.

Statbel utilise le logiciel open source et le langage de programmation R pour effectuer le *webscraping*. Ce logiciel contient des packages qui permettent d'effectuer le *webscraping*: *rvest*, qui permet de collecter des données de pages internet statiques, et *RSelenium* et *phantomjs*, pour les interactions dynamiques avec les pages internet (par exemple cliquer sur un bouton, remplir des formulaires, faire défiler indéfiniment, …). De plus, avec CSS Selectors[6] ou XPath, il est possible de rechercher les informations requises dans la structure complexe d'une page html. Ci-dessous un exemple de CSS Selector:

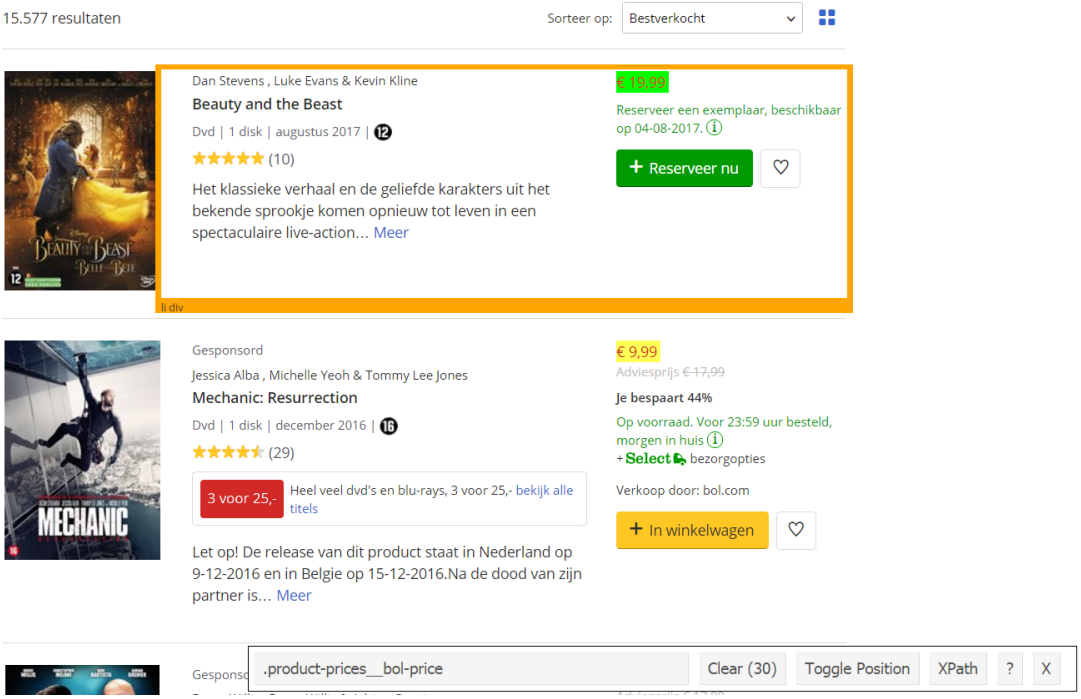


Figure 1: Exemple de CSS Selector

Le CSS Selector pour le prix du produit est '. product prices__bol-price'.

Le code programmé pour visiter le site internet et extraire les données est enregistré pour chaque site internet dans un script. Les données qui font généralement l'objet d'un *scraping* et sont donc collectées: le titre du produit, la description du produit, la (sous-)catégorie, l'ID et le prix du produit (prix en promotion et hors promotion). Techniquement, il s'agit de programmer un script via R pour visiter virtuellement un site web, de collecter toutes les informations intéressantes du site internet et de les stocker dans ce que l'on appelle un *dataframe*.

Les scripts/robots sont exécutés via un serveur et s'identifient toujours en tant que 'Statistics Belgium'. Afin de ne pas surcharger les sites internet, certains scripts sont exécutés la nuit et il est possible de programmer quand les scripts doivent tourner (par exemple, chaque jour, 2 X par semaine, chaque semaine,...).

Contrairement à la collecte manuelle de données, lors de laquelle les prix sont collectés selon des définitions spécifiques, le *webscraping* collecte le plus possible de données. Ces données peuvent ensuite être traitées (éventuellement par apprentissage automatique) afin de calculer des indices des prix.

[6] CSS fait référence aux Cascading Style Sheets, qui déterminent la mise en page de chaque élément d'une page Web. L'élément en lui-même est défini au travers de l'utilisation d'une balise HTML. Les CSS Selectors sont utilisés pour sélectionner les éléments HTML.

Il est en effet possible qu'un site internet ne contienne pas la même quantité de données pour tous les éléments. Ainsi, le produit 1 peut être affiché avec uniquement un prix original mais le produit 2 avec un prix original et un prix promotionnel (voir figure 2). Avec le sélecteur CSS, on voit que la balise du prix du produit 1 est '.current-price'. Pour le produit 2, cette balise correspond au prix promotionnel.

Figure 2: Exemple de produits avec des nombres de prix différents. Les prix sélectionnés ont comme balise 'current.price'.

Le prix original du produit 2 a comme balise: '.previous-price'.

Figure 3: Le prix sélectionné (vert) a comme balise '.previous-price'.

Il est conseillé de collecter tous les prix, 1 prix promotionnel et 2 prix non promotionnels. Pour le produit 1, il faut toutefois indiquer qu'il n'y a pas de prix promotionnel. A cet effet, une fonction de *scraping* a été développée qui indique une caractéristique comme NA (*not available*) si la caractéristique n'a pas de valeur:

```
scrape_css <- function(css,group){
                txt <- main_page %>%
                html_nodes(group) %>%
                lapply(. %>% html_nodes(css) %>% html_text() %>%
                ifelse(identical(., character(0)), NA, .)) %>%
                unlist
                return(txt)
                }
```

De manière générale, un script est construit de la manière suivante:

- Chargement de la page d'accueil du site internet
- Contrôle et scraping de différentes catégories et sous-catégories
- Parcourir les catégories et effectuer le scraping des caractéristiques.
  - Prix
  - ID/sku
  - Description du produit
  - Url du produit
- Rassembler les données dans une dataframe (= tableau)

Pour les sites internet de voyage, les billets d'avion ou les véhicules d'occasion, par exemple, il est plus commode de définir un certain nombre de critères à l'avance. Le script ne parcourra alors que les éléments correspondants à ces critères. Ainsi, les mêmes articles sont recherchés chaque fois et les résultats peuvent être agrégés et comparés correctement. Souvent, sur ces sites internet, ces critères doivent être introduits dans une "fonction de recherche". Le script est donc programmé de manière à imiter une telle action. Il s'agit alors de ce que l'on appelle une réservation virtuelle.

Il est également possible d'effectuer un *scraping* de certaines caractéristiques d'un produit, par exemple dans le domaine des produits électroniques grand public. Sur cette base, on peut calculer des indices des prix hédoniques (voir plus loin).

Les produits présentés sur un site internet sont souvent répartis sur différentes pages. Le script développé doit donc aussi en tenir compte et passer automatiquement à une page suivante lorsque le dernier élément de la page précédente est lu. A cet effet, on peut programmer de telle manière que le script clique sur le bouton *'suivant'* comme le ferait un utilisateur ordinaire. Cependant, il existe aussi des sites internet qui chargent plus de produits à mesure que le visiteur déroule la page vers le bas. Ce type de déroulement peut également être simulé par le script, afin que tous les produits disponibles puissent faire l'objet d'un *scraping*.

Le *webscraping* permet de collecter en continu des données en ligne. D'un point de vue quantitatif, toutes les données disponibles sur les sites internet font l'objet d'un *scraping* (*bulk scraping*). Les données peuvent également être recherchées de manière ciblée sur la base de critères prédéterminés, comme cela se ferait manuellement (cf. outil robotisé, chapitre 12).

Actuellement, une soixantaine de scripts de webscraping sont exécutés notamment pour les segments suivants:

- Habillement
- Chaussures
- Hôtels
- Billets d'avion
- Billets de trains internationaux
- Véhicules d'occasion
- Droguerie
- Livres
- DVD & disques Blu-ray
- Jeux vidéo
- Produits électroniques grand public
- Chambres d'étudiants
- Supermarchés
- …

Les chapitres suivants présentent des résultats expérimentaux pour certains segments. Comme il s'agit encore de résultats expérimentaux, l'indication de la période dans les graphiques a été remplacée par un chiffre au lieu du mois réel. Il s'agit d'une procédure standard dans les statistiques de prix dans le cadre de recherches afin d'éviter le recouplage avec les indices officiels, qui pourrait aboutir à d'éventuelles interprétations erronées. Normalement, la période se réfère au mois. Dans le cas contraire, le texte qui explique le graphique le mentionne.

5. DE CAS - PRODUITS ÉLECTRONIQUES GRAND PUBLIC

Les produits électroniques grand public sont un domaine pour lequel de nombreuses données sont disponibles en ligne. Statbel a effectué un *scraping* de données pendant au moins 17 mois pour divers appareils électroniques, notamment les ordinateurs portables, les tablettes, les réfrigérateurs et les machines à laver. Ces données ont été utilisées pour examiner différentes méthodes de calcul des indices.

Il est difficile de mesurer l'évolution des prix des produits électroniques grand public en raison de trois facteurs:

1. Les produits ne sont sur le marché que pour une courte période de temps. Il y a donc un important flux entrant de nouveaux produits, qui s'accompagne d'un flux sortant simultané d'anciens produits.
2. Ces produits plus anciens quittent le marché à un prix moins élevé (dumping ou soldes) que ceux auxquels ils sont arrivés sur le marché.
3. Les produits qui arrivent sur le marché ont souvent des caractéristiques différentes - ils sont de meilleure qualité (par exemple, plus économes en énergie ou, dans le cas des ordinateurs, plus rapides) - que les produits qui disparaissent de la gamme.

Ces facteurs entraînent un certain nombre de problèmes :

1. Le premier point a comme conséquence qu'il n'est pas possible de mesurer l'évolution des prix d'un même produit sur une longue période. Des remplacements sont donc souvent nécessaires.
2. Cela nous amène au deuxième point, à savoir la baisse du prix d'un produit électronique au cours de sa durée de vie. Ce point a comme conséquence que si l'on mesure l'évolution des prix des mêmes produits sur la base d'un indice de Jevons en chaîne mensuel et si les produits restent dans l'échantillon jusqu'à ce qu'ils quittent l'offre, cela entrainera une dérive en chaîne négative (*downward chaindrift*). Le graphique suivant montre un indice en chaîne mensuel et un indice en chaîne quotidien pour les ordinateurs portables:

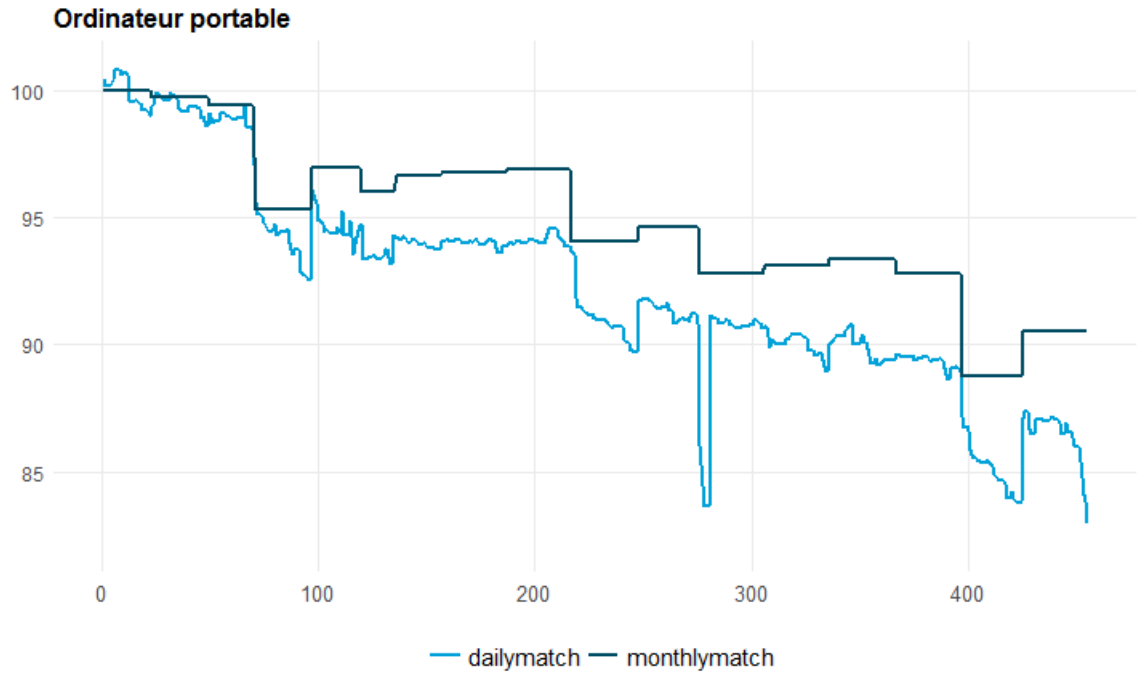


Figure 4: Comparaison entre un indice en chaîne quotidien et mensuel

Dans l'indice en chaîne, les nouveaux produits sont couplés à l'évolution des prix mesurée jusqu'à ce moment. Ce couplage neutralise donc la différence de prix entre l'ancien et le nouveau produit et la prend implicitement en compte via l'évolution des prix des autres articles correspondants au cours des deux mois adjacents.

L'indice calculé ci-dessus est donc un indice calculé à l'aide de produits disponibles pendant deux périodes adjacentes (*matched model index*). La formule de l'indice de Jevons est donc la suivante:

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

où $G_{0,1}$ est l'échantillon. Il s'agit du set de tous les produits qui correspondent pendant deux périodes successives (période 0 et période 1). Dans un indice en chaîne de plusieurs périodes, ces indices mensuels sont couplés en multipliant les indices obtenus entre eux. Pour un indice de Jevons en chaîne comprenant trois périodes, la formule est la suivante:

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

Le graphique montre que l'indice en chaîne mensuel diminue de manière moins prononcée. La raison en est que, dans une agrégation mensuelle, le prix final du produit est en partie lissé par les prix plus élevés en vigueur pendant le reste du mois. Si le prix final inférieur d'un produit était introduit chaque fois au début du mois et si le produit quittait la gamme à la fin du mois, les indices en chaîne quotidien et mensuel seraient alors logiquement identiques. La collecte des données classique tente d'éviter le problème de la dérive en chaîne négative (*downward chaindrift*) décrit ci-dessus en travaillant avec le *Monthly Chaining and Re-sampling* (voir ci-dessous).

Une explication théorique des différents modèles d'indices est présentée ci-dessous, dans lesquels les modèles sont appliqués aux données collectées par *webscraping*. Une partie des modèles développés ci-dessous, seront aussi utilisés dans d'autres études de cas.

5.1. GEKS & RYGEKS

Outre le couplage de deux mois consécutifs, il existe également des *matched model indices* plus avancés tels que l'indice GEKS (nommé d'après Gini, Eltetö, Köves & Szulc) dans lequel non seulement la dernière et l'avant-dernière période sont chaînées l'une avec l'autre, mais aussi toutes les périodes intermédiaires dans lesquelles les produits correspondent. Les indices GEKS ont été développés pour les comparaisons spatiales de prix (ils notamment sont utilisés pour les parités de pouvoir d'achat). L'objectif de la procédure GEKS traditionnelle est que lorsque le pouvoir d'achat du pays A est directement comparé à celui du pays C, la comparaison devrait produire le même résultat que si cette comparaison était effectuée indirectement en comparant d'abord le pays A avec le pays B, puis le pays B avec le pays C. La procédure GEKS est donc transitive en ce sens qu'elle est indépendante du pays que l'on choisit comme base. Cette méthode GEKS n'est donc logiquement pas utilisée pour les comparaisons spatiales des prix pour les indices des prix à la consommation, mais pour les comparaisons de prix au fil du temps (temporelles).

La méthode utilise tous les produits qui correspondent les uns aux autres en appliquant ensuite une moyenne géométrique non pondérée (indice de Jevons)[7] de tous les indices de prix bilatéraux possibles, chaque période servant de période de référence. Cette moyenne n'est pas pondérée car aucune information de pondération n'est disponible. Un indice de Jevons GEKS (ou GEKS-J) pour la période de 0 à t peut être exprimé de la manière suivante:

$$GEKS_J^{0,t} = \prod_{l=0}^{T} \left(P^{0l} \Big/ P^{tl}\right)^{(1/T+1)} = \prod_{l=0}^{T} (P^{0l} P^{lt})^{(1/T+1)} \qquad \text{Équ. 3}$$

où P^{0l} et P^{lt} sont des indices de Jevons bilatéraux entre respectivement la période 0 et la période l (avec 0 comme période de base) et la période l et la période t (avec l comme période de référence) et $P^{lt} = 1/P^{tl}$. Où 0, l et t sont des périodes de temps de la série chronologique T (=0, 1, …, T). Dans les indices des prix bilatéraux, seuls les éléments qui correspondent (*matched items*) sont utilisés. Dans formule ci-dessus, l'indice des prix de la période t dépend aussi des prix des périodes

[7] Pour plus d'informations sur ce choix : Roels, D., Van Loon, K. (2017) *L'utilisation des données de scanning des supermarchés dans l'indice des prix à la consommation*

futures qui ne sont pas encore disponibles (t+1, t+2 , …). Dans la pratique, on ne recalcule que la période la plus récente T. La formule peut dès lors être réécrite comme suit:

$$GEKS_J^{0,T} = \prod_{t=0}^{T} (P^{0t} P^{tT})^{(1/T+1)} \qquad \text{Équ. 4}$$

Si, dans le cas d'un indice GEKS-J, la période est étendue (T+1), par exemple au mois prochain pour l'IPC, alors la comparaison prend la forme suivante:

$$GEKS_J^{0,T+1} = \prod_{t=0}^{T+1} (P^{0t} P^{tT+1})^{(1/T+2)} \qquad \text{Équ. 5}$$

Comme tous les indices bilatéraux de l'indice GEKS-J sont calculés sur la base de toutes les correspondances dans le dataset, les indices pour les périodes 1,2,…, T - calculés au moment T+1 sur la base de l'équation ci-dessus - sont donc différents de ceux calculés précédemment. Cela signifierait dès lors que les indices antérieurs devraient être "revus" à chaque fois, ce qui est évidemment problématique pour les indices déjà publiés. Pour résoudre ce problème de révisions, on peut travailler une année mobile (*rolling year*) en utilisant une période mobile de 13 mois pour calculer les indices. Le GEKS-J est alors appelé RYGEKS-J. Ainsi, les résultats ne doivent pas être revus. On utilise 13 mois car cela correspond à un an (par exemple, de décembre à décembre) dans le calcul de l'inflation et représente donc la période la plus courte possible pour les produits soumis à des variations saisonnières. Dans le RYGEKS-J, l'indice GEKS-J mensuel le plus récent est couplé aux séries chronologiques existantes ayant un indice GEKS-J des 13 premiers mois comme point de départ. Cela signifie que le point de départ du RYGEKS-J est égal à:

$$GEKS_J^{0,12} = \prod_{t=0}^{12} (P^{0t} P^{t12})^{(1/13)} \qquad \text{Équ. 6}$$

L'indice GEKS-J obtenu peut maintenant être multiplié par l'indice GEKS-J mensuel le plus récent pour obtenir l'indice du mois en cours. Le mois suivant, l'indice qui vient d'être obtenu peut à nouveau être multiplié par le nouvel indice mensuel. La formule est la suivante[8]:

$$\begin{aligned}
RYGEKS_J^{0,13} &= GEKS_{0,12}^{0,12} * GEKS_{1,13}^{12,13} \\
&= GEKS_J^{0,12} \prod_{t=1}^{13} (P^{12t} P^{t13})^{(1/13)} \\
RYGEKS_J^{0,14} &= RYGEKS_J^{0,13} \prod_{t=2}^{14} (P^{13t} P^{t14})^{(1/13)} \qquad \text{Équ. 7} \\
RYGEKS_J^{0,15} &= \cdots
\end{aligned}$$

La figure suivante présente une comparaison entre GEKS et RYGEKS. Les deux modèles ont été appliqués aux données sur les produits électroniques obtenues par *scraping*. Une très faible différence peut être observée à partir du 14e mois. Comme expliqué ci-dessus, il n'y a en effet pas de différence les 13 premiers mois.

[8] La notation $I_{c,d}^{a,b}$ représente (la méthode de) l'indice I pour la période a à b avec une fenêtre sur la période c à d. La fenêtre indique la période pendant laquelle les données sont utilisées, voir également la section 5.2.2. Lorsqu'un indice$J$ est utilisé, celui-ci renvoie à l'utilisation d'un indice de Jevons au niveau élémentaire. Cet indice est mis de côté dans le reste du texte.

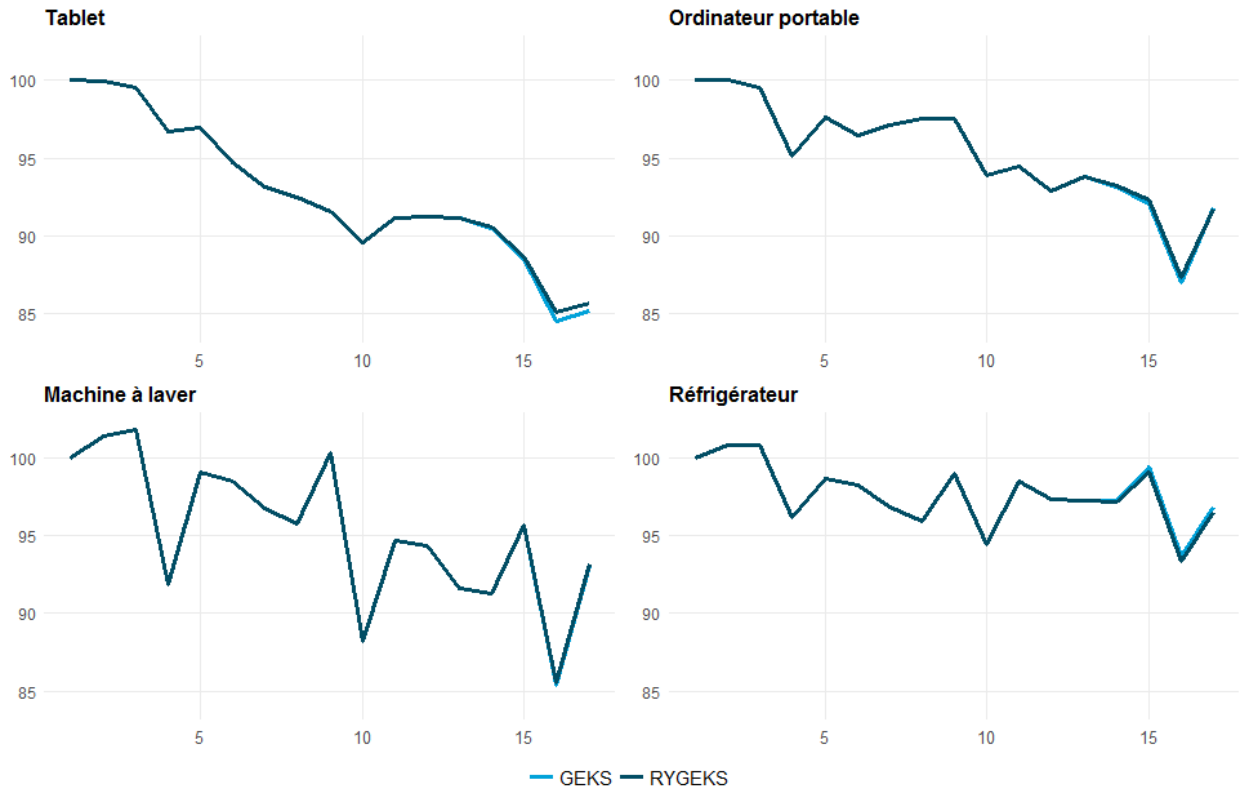


Figure 5: Comparaison entre le GEKS et le RYGEKS

De plus, le fait de travailler avec un *matched model index* multilatéral tel que le GEKS-J ne résout pas le problème de dérive. Il n'y a pas de rebond après 13 mois. En raison de cette dérive, la différence de prix entre les produits qui sortent de la gamme et ceux qui font leur apparition dans la gamme doit être prise en compte d'une manière ou d'une autre dans le calcul des indices. L'intégration directe de cet écart de prix dans le calcul entrainerait un biais de l'inflation à la hausse si les nouveaux produits étaient plus chers que les anciens (voir graphique ci-dessous). La différence de qualité ne serait pas non plus prise en compte. Les produits incomparables seraient comparés entre eux, ce qui n'est logiquement pas conforme à la réglementation de l'indice des prix à la consommation harmonisé (IPCH) car ceci est contraire à l'un des principes de base de la mesure de l'inflation.

Figure 6: comparaison entre *average price* et *monthly matching*

Eurostat propose (également dans le futur manuel de l'IPCH) d'utiliser une variante de l'indice en chaîne mensuel - *Monthly Chaining and Re-sampling* (MCR) - si l'on décide malgré tout de travailler avec un indice en chaîne mensuel, étant donné que l'indice en chaîne mensuel reste la méthode la plus couramment utilisée pour les relevés de prix classiques (elle est également

appliquée en Belgique pour les relevés de prix classiques des produits électroniques grand public). Dans cette variante, les anciens produits doivent être remplacés quand ils ont encore un prix représentatif et les nouveaux produits doivent également être repris dans l'échantillon à un prix représentatif. Le nouveau produit est également couplé à l'indice qui existe jusqu'alors. Cela signifie que la différence de prix entre l'ancien et le nouveau produit n'est pas directement intégrée dans le calcul, mais est implicitement captée par l'évolution des prix des produits inclus dans l'échantillon au cours des deux mois consécutifs.

Il n'est pas tout à fait surprenant que cela soit difficile à mettre en pratique. Il est en effet difficile de déterminer quand le prix d'un produit est représentatif. Le fait qu'il en résulte des taux d'inflation différents pour les produits électroniques grand public sur la base de la même méthodologie n'est pas surprenant non plus, non seulement parce que le moment du remplacement peut avoir un impact, mais aussi parce que la politique de fixation des prix des magasins exerce une influence sur l'évolution des prix mesurée. Si les produits sont introduits à un prix faible (promotion de lancement) qui augmente ensuite légèrement puis diminue à nouveau, cela donnera une évolution de prix différente que si un produit ne connait qu'une baisse de prix après son prix de lancement. De même, une politique de baisse graduelle des prix au cours de la durée de vie donnera des taux d'inflation différents d'une politique qui ne prévoit que peu de baisses des prix pendant la durée de vie et une forte baisse en fin de vie du produit (où il convient donc de le remplacer et de ne pas tenir compte de cette baisse en raison de la *chain drift* mentionnée ci-dessus).

5.2. Régression hédonique

L'un des avantages potentiels du *webscraping* est qu'il est également possible d'enregistrer les caractéristiques des produits, ce qui permet également d'utiliser des méthodes plus explicites pour intégrer les changements de qualité dans le calcul (méthodes hédoniques). Il est difficile d'appliquer ces méthodes aux relevés de prix classiques (en ligne ou hors ligne). Noter toutes les caractéristiques est une tâche intensive. De plus, lors d'un relevé de prix classique, il ne suffit pas seulement d'enregistrer les caractéristiques des produits issus de l'échantillon. Cela génère trop peu de résultats pour pouvoir appliquer des techniques de régression correctes.

Les techniques de régression des indices des prix à la consommation sont appelées régressions hédoniques. Le prix est toujours exprimé en fonction des caractéristiques.

Lors du remplacement d'un ancien produit par un nouveau, les méthodes hédoniques n'ont pas pour objectif d'estimer la différence de qualité sur la base du modèle de régression (après quoi le prix peut être adapté pour tenir compte de cette différence). Cependant, elles estiment quel serait le prix relatif (ou le prix) d'un nouveau produit au cours de la période de référence ou d'un produit disparu au cours de la période actuelle.

En outre, les ajustements de qualité sont pris en compte dans l'ensemble de la population, et non seulement dans l'échantillon des produits pour lesquels des remplacements ont eu lieu. Ainsi, les remplacements 1 pour 1 ne seraient pas possibles, par exemple parce que le nombre de produits dans un segment diminuerait au fil du temps, et qu'il n'y aurait pas de nouveau remplaçant pour chaque produit disparu.

La forme fonctionnelle de régression hédonique la plus couramment utilisée est un modèle semi-logarithmique linéaire dans lequel le prix subit une transformation logarithmique[9]. Cette transformation logarithmique fait notamment en sorte que les résidus du modèle sont moins caractérisés par l'hétéroscédasticité. La forme fonctionnelle peut être exprimée comme suit pour une série chronologique avec deux périodes adjacentes (t=0,1):

$$\ln(p_i^t) = \alpha^t + \sum_{k=1}^{K} \beta_k^t x_{ik} + \varepsilon_i^t \qquad \textit{Équ. 8}$$

Où p_i^t est le prix du produit i pendant la période t, dont on prend ensuite le logarithme naturel. x_{ik} représente la caractéristique k (k=1, …, K) du produit i, avec le paramètre correspondant β_k^t. α^t est l'intercept, les termes de perturbation ε_i^t sont supposés indépendamment avec une valeur probable de 0 et une variance constante.

[9] Pour plus d'informations sur la régression hédonique : Triplet, J. (2006), *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*. Paris: OCDE. de Haan, J. (2010) '*Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and 'Re-Pricing' Methods*', Journal of Economics and Statistics Vol. 230, No. 6, Themenheft: Index Number Theory and Price Statistics pp. 772-791.

5.2.1. Time Dummy Hedonics

La variante la plus couramment utilisée de la forme fonctionnelle décrite ci-dessus est une méthode hédonique avec *time dummy* dans laquelle l'équation prend la forme suivante pour le même contexte avec deux périodes adjacentes (t=0,1):

$$\ln(p_i^t) = \alpha + \delta D_i + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i^t \qquad \textit{Équ. 9}$$

Avec D_i comme variable *time dummy* prenant la valeur 1 si le produit i est disponible pendant la période 1 et 0 s'il n'est pas disponible. Les valeurs pour le terme constant α et les paramètres des caractéristiques β_k sont supposés stables pour les deux périodes ($\beta_k^0 = \beta_k^1 = \beta_k$). Ici aussi, on suppose que les termes de perturbation ε_i^t sont distribués indépendamment avec une espérance nulle et une variance égale. Le paramètre *time dummy* δ mesure l'effet du temps sur le logarithme naturel du prix, bien entendu en tenant compte des caractéristiques déterminant la qualité. L'indice *time dummy* $TD_{0,1}^{0,1}$ peut alors être obtenu en prenant l'exponentielle de δ:

$$TD_{0,1}^{0,1} = \exp(\hat{\delta}) \qquad \textit{Équ. 10}$$

Cet indice peut donc être calculé directement à partir de l'équation, de sorte que la méthode hédonique *time dummy* est souvent appelée méthode hédonique directe. Les prix estimés de la période 0 ($\hat{p}_i^0$) et de la période 1 ($\hat{p}_i^1$) peuvent être obtenus à partir de l'équation 9.

$$\hat{p}_i^0 = exp\left(\hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k x_{ik}\right) \qquad \textit{Équ. 11}$$

$$\hat{p}_i^1 = exp\left(\hat{\alpha} + \hat{\delta} + \sum_{k=1}^{K} \hat{\beta}_k x_{ik}\right) \qquad \textit{Équ. 12}$$

Cela montre donc que l'indice *time dummy* $TD_{0,1}^{0,1}$ est égal au rapport entre les prix estimés de tous les articles pour la période 1 ($\hat{p}_i^1$) et ceux de tous les articles pour la période 0 ($\hat{p}_i^0$), puisque l'équation 10, qui peut être réécrite à partir des équations 11 et 12, est pour chaque produit i:

$$TD_{0,1}^{0,1} = \exp(\hat{\delta}) = \frac{exp\left(\hat{\alpha} + \hat{\delta} + \sum_{k=1}^{K} \hat{\beta}_k x_{ik}\right)}{exp\left(\hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k x_{ik}\right)} = \frac{\hat{p}_i^1}{\hat{p}_i^0} \qquad \textit{Équ. 13}$$

Étant donné que la somme des résidus est égale à zéro pour les deux périodes, compte tenu de l'inclusion du terme constant (et également de la variable *time dummy*), la différence entre la somme des prix effectifs et estimés est logiquement égale à zéro pour un ensemble G_0 de produits disponibles i pendant la période 0 et un set G_1 de produits disponibles i au cours de la période 1.

$$\sum_{i \in G_0} (ln\, p_i^0 - ln\, \hat{p}_i^0) = \sum_{i \in G_1} (ln\, p_i^1 - ln\, \hat{p}_i^1) = 0 \qquad \textit{Équ. 14}$$

Ce que l'on peut réécrire simplement comme:

$$\sum_{i \in G_0} ln\left(\frac{p_i^0}{\hat{p}_i^0}\right) = \sum_{i \in G_0} ln\left(\frac{p_i^1}{\hat{p}_i^1}\right) = 0 \qquad \textit{Équ. 15}$$

Si l'on prend maintenant l'exponentielle, on obtient alors:

$$\exp(0) = 1 = \prod_{i \in G_0} \left(\frac{p_i^0}{\hat{p}_i^0}\right) = \prod_{i \in G_1} \left(\frac{p_i^1}{\hat{p}_i^1}\right) \qquad \textit{Équ. 16}$$

Dans le cas où $G_0 = G_1 = G$ l'équation 16 est:

$$\prod_{i \in G}\left(\frac{p_i^0}{\hat{p}_i^0}\right) = \prod_{i \in G}\left(\frac{p_i^1}{\hat{p}_i^1}\right) \qquad \textit{Équ. 17}$$

Ce que l'on peut réécrire comme:

$$\prod_{i \in G}\left(\frac{\hat{p}_i^1}{\hat{p}_i^0}\right) = \prod_{i \in G}\left(\frac{p_i^1}{p_i^0}\right) \qquad \textit{Équ. 18}$$

Par la suite, s'applique également à G avec N produits:

$$\prod_{i \in G}\left(\frac{\hat{p}_i^1}{\hat{p}_i^0}\right)^{1/N} = \prod_{i \in G}\left(\frac{p_i^1}{p_i^0}\right)^{1/N} = P_J^{0,1} \qquad \textit{Équ. 19}$$

ce qui correspond à l'indice de Jevons déjà décrit ci-dessus dans l'équation 1 ($\prod_{i \in G}\left(\frac{p_i^1}{p_i^0}\right)^{1/N} = P_J^{0,1}$).

Puisqu'il découle de l'équation 13 que pour tous les produits $i$le ratio est $\frac{\hat{p}_i^1}{\hat{p}_i^0} = TD_{0,1}^{0,1}$ alors logiquement:

$$TD_{0,1}^{0,1} = \prod_{i \in G}\left(\frac{\hat{p}_i^1}{\hat{p}_i^0}\right)^{1/N} = \prod_{i \in G}\left(\frac{p_i^1}{p_i^0}\right)^{1/N} = P_J^{0,1} \qquad \textit{Équ. 20}$$

Par conséquent, la méthode hédonique *time dummy* dans la variante semi-logarithmique correspond à une formule d'indice élémentaire basée sur la moyenne géométrique qui est également utilisée dans la méthodologie classique.

Dans le cas où $G_0 \neq G_1$, qui se est normalement le cas le plus fréquet, l'indice est calculé avec des articles qui correspondent entre les deux périodes (conformément au *matched Jevons index*, nous décrivons ce set comme $G_{0,1}$), les articles qui ont disparu de la période zéro (G_0^D) et les nouveaux articles de la période 1 (G_1^N). Le set complet d'articles de la période 0 est alors $G_0 = G_{0,1} \cup G_0^D$et celui de la période 1 est = G_1= ($G_{0,1} \cup G_1^N$). Il découle des équations 11 et 12, en prenant la moyenne géométrique de respectivement G_0 avec N_0 produits et G_1avec N_1produits:

$$\prod_{i \in G_0}(\hat{p}_i^0)^{1/N_0} = exp(\hat{\alpha})exp\left(\sum_{k=1}^{K}\hat{\beta}_k \sum_{i \in G_0} x_{ik}/N_0\right) \qquad \textit{Équ. 21}$$

$$\prod_{i \in G_1}(\hat{p}_i^1)^{1/N_1} = exp(\hat{\alpha})\, exp\left(\hat{\delta}\right)exp\left(\sum_{k=1}^{K}\hat{\beta}_k \sum_{i \in G_1} x_{ik}/N_1\right) \qquad \textit{Équ. 22}$$

Par conséquent, en divisant l'équation 22 par l'équation 21, l'indice *time dummy* est égal à:

$$\begin{aligned} TD_{0,1}^{0,1} &= exp\left(\hat{\delta}\right) \\ &= \frac{\prod_{i \in G_1}(\hat{p}_i^1)^{1/N_1} exp\left(\sum_{k=1}^{K}\hat{\beta}_k \sum_{i \in G_0} x_{ik}/N_0\right)}{\prod_{i \in G_0}(\hat{p}_i^0)^{1/N_0}\, exp\left(\sum_{k=1}^{K}\hat{\beta}_k \sum_{i \in G_1} x_{ik}/N_1\right)} \\ &= \frac{\prod_{i \in G_1}(\hat{p}_i^1)^{1/N_1}}{\prod_{i \in G_0}(\hat{p}_i^0)^{1/N_0}}\, exp\left(\sum_{k=1}^{K}\hat{\beta}_k(\bar{x}_k^0 - \bar{x}_k^1)\right), \end{aligned} \qquad \textit{Équ. 23}$$

où $\bar{x}_k^0 = \sum_{i \in G_0} x_{ik}/N_0$ et $\bar{x}_k^1 = \sum_{i \in G_1} x_{ik}/N_1$ et sont donc tous les deux égaux à la moyenne de la caractéristique k de la période 0 et de la période 1. Etant donné qu'en raison de l'intégration de la variable *time dummy* et du terme constant, la somme des résidus est ici aussi égale à zéro pendant chaque période, il ressort de l'équation 16 que $\prod_{i \in G_0}(\hat{p}_i^0)^{1/N_0} = \prod_{i \in G_0}(p_i^0)^{1/N_0}$ et $\prod_{i \in G_1}(\hat{p}_i^1)^{1/N_1} = \prod_{i \in G_1}(p_i^1)^{1/N_1}$. Dès lors, l'équation 23 peut aussi être réécrite comme suit

$$TD_{0,1}^{0,1} = exp(\hat{\delta}) = \frac{\prod_{i \in G_1}(p_i^1)^{1/N_1}}{\prod_{i \in G_0}(p_i^0)^{1/N_0}} \, exp\left(\sum_{k=1}^{K} \hat{\beta}_k(\bar{x}_k^0 - \bar{x}_k^1)\right) \qquad \textit{Équ. 24}$$

L'indice *time dummy* est donc égal au ratio de la moyenne géométrique des prix (un *unmatched Jevons index*) qui est ajusté via un facteur qui effectue un ajustement pour la qualité ($exp\left(\sum_{k=1}^{K} \hat{\beta}_k(\bar{x}_k^0 - \bar{x}_k^1)\right)$) sur la base des modifications des caractéristiques entre la période 0 et la période 1.

Dans le cas où les caractéristiques moyennes entre les deux périodes sont identiques ($\bar{x}_k^0 = \bar{x}_k^1$), ce qui peut se produire si l'échantillon d'articles est identique ou si un nouveau produit a les mêmes caractéristiques qu'un produit disparu, alors ce facteur est égal à 1 et l'indice est donc calculé comme un simple ratio de la moyenne géométrique des prix et correspond donc soit à un *matched Jevons index* s'il n'y a pas eu de changement dans l'échantillon ($G_0 = G_1$), soit à un *unmatched Jevons index*, s'il y a eu des changements dans l'échantillon ($G_0 \neq G_1$).

Dans la pratique, un indice n'est pas calculé sur deux périodes comme dans les équations ci-dessus, mais sur plusieurs périodes t= 0,..., T de sorte que l'équation 9 prend la forme suivante:

$$\ln(p_i^{0,T}) = \hat{\alpha} + \sum_{t=1}^{T} \hat{\delta}^t {D_i}^t + \sum_{k=1}^{K} \hat{\beta}_k x_{ik} \qquad \textit{Équ. 25}$$

où ${D_i}^t$est la variable *time dummy* qui prend la valeur 1 si le produit i est disponible pendant la période t et 0 s'il n'est pas disponible. L'antilogarithme (l'exponentielle) de δ^t donne l'indice *time dummy* ($TD_{0,T}^{0,t}$) pour la période t $(t = 0,1,\ldots,T)$. Dans un contexte de périodes multiples, les paramètres sont donc maintenus constants tout au long de la période. L'inconvénient de cet indice à périodes multiples est qu'il est sujet à des révisions chaque fois que la période est prolongée. Pour la période T+1, le modèle de régression sera réévalué et les indices pour la période T seront donc révisés.

La figure suivante montre les résultats de la méthode TD par rapport à la méthode GEKS. La sélection des caractéristiques incluses dans l'équation de régression a été déterminée sur la base de l'analyse du niveau de signification des paramètres, le *critère d'information d'Akaike (AIC)* et le *critère d'information bayésien (BIC)*[10]. On a également examiné la logique du rapport entre la caractéristique et le prix. Par exemple, le coefficient du poids d'un produit dans l'équation de régression sera la plupart du temps positif, mais les consommateurs préfèrent toutes autres choses égales par ailleurs, les produits plus légers aux produits plus lourds. Ces variables n'ont donc logiquement pas été incluses dans l'équation de régression.

[10] L'AIC est une mesure de la qualité relative d'un modèle statistique pour un set de données déterminé. Le BIC est un critère pour sélectionner un modèle déterminé parmi une sélection limitée de modèles. Les deux critères sont liés l'un à l'autre.

Figure 7: Comparaison entre la méthode *Pooled Time Dummy* et la méthode GEKS

Les deux indices montrent une évolution similaire de baisse des prix. Pour trois des quatre produits, l'indice *time dummy* est supérieur à l'indice GEKS et il y aurait donc une dérive (*drift*) si l'on utilisait la méthode GEKS. Ne pas traiter la différence de prix entre les produits qui sortent de la gamme et ceux qui sont nouveaux dans la gamme entraîne une surestimation de la baisse des prix. Si l'on en tient compte au travers de la régression hédonique, la baisse des prix est moins prononcée. L'indice des prix hédonique (TD) estimé dans le segment des réfrigérateurs dépasse l'indice GEKS, mais il s'agit d'une image biaisée. Il y a deux raisons à cela: d'une part, le taux d'attrition (rotation des articles) des réfrigérateurs n'est pas très élevé. Le graphique suivant le montre bien:

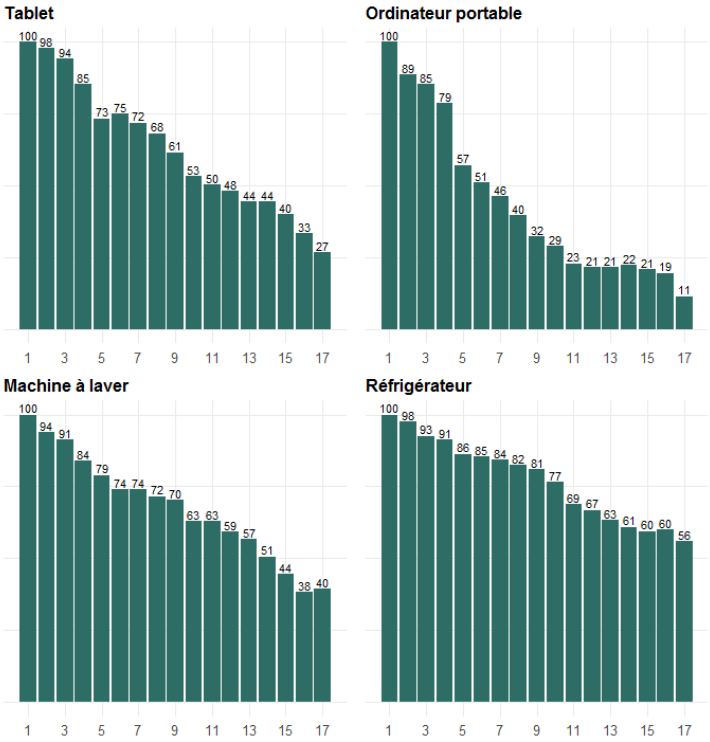


Figure 8: taux d'attrition des produits électroniques grand public sélectionnés

Le nombre d'articles qui correspondent après 17 mois est toujours de 56 %, contrairement aux tablettes (27 %) ou aux ordinateurs portables (11 %). La différence de prix entre les produits qui sortent de la gamme et les nouveaux produits de la gamme jouera donc un rôle moins décisif au cours de la période considérée. D'autre part, pendant le mois choisi comme point de départ, des produits sont apparemment plus chers que ce que l'on aurait pu prévoir sur la base de leurs caractéristiques, ce qui fait grimper le prix hédonique. Cela apparait clairement lorsque le nombre de périodes est augmenté, comme on peut le voir sur la figure 12 (cf. 5.2.4 Indice *Time Dummy* avec *Window Splice*). Une augmentation peut être observée à partir de la période 8, qui a été prise comme base dans les résultats précédents, neutralisant ainsi la différence. L'évolution part ensuite de cette base, ce qui aboutit à la figure 7.

5.2.2. Indice *Time Dummy* avec période mobile (*Rolling Window*)

Par analogie avec le RYGEKS précité, une période mobile (*rolling window*) peut également être utilisée pour un indice *time dummy*, dans lequel la durée de la période utilisée reste la même, à savoir 13 mois ($T = 0, \dots, 12$). Cet indice avec période mobile (*rolling year time dummy* - RYTD) est dès lors égal à:

$$RYTD_{0,12}^{0,12} = \exp(\hat{\delta}^{12}) \qquad \textit{Équ. 26}$$

Avec modèle initiale $\hat{\alpha} + \sum_{t=1}^{12} \hat{\delta}^t {D_i}^t + \sum_{k=1}^{K} \hat{\beta}_k x_{ik}$. Le modèle pour le mois suivant (T+1) est ensuite estimé sur la base des périodes 1 à 13 (au lieu de 0 à 12):

$$\hat{\alpha} + \sum_{t=2}^{13} \hat{\delta}^t {D_i}^t + \sum_{k=1}^{K} \hat{\beta}_k x_{ik} \qquad \textit{Équ. 27}$$

L'indice pour la période de 0 à 13 peut ensuite être calculé en couplant la plus récente évolution des prix d'une période à l'autre, soit celle entre le 13e ($RYTD_{1,13}^{1,13}$) et le 12e ($RYTD_{1,13}^{1,12}$) mois, aux séries chronologiques obtenues précédemment jusqu'au 12e mois:

$$\begin{aligned} RYTD^{0,13} &= RYTD_{0,12}^{0,12} * \left({RYTD_{1,13}^{1,13}} \Big/ {RYTD_{1,13}^{1,12}} \right) \qquad \textit{Équ. 28} \\ &= RYTD_{0,12}^{0,12} * RYTD_{1,13}^{12,13} \end{aligned}$$

Le principe pour les prochains mois est logiquement identique:

$$\begin{aligned} RYTD^{0,14} &= RYTD^{0,13} * \left({RYTD_{2,14}^{2,14}} \Big/ {RYTD_{2,14}^{2,13}} \right) \qquad \textit{Équ. 29} \\ RYTD^{0,15} &= \cdots \end{aligned}$$

Étant donné que cette méthode couple la dernière évolution d'une période à l'autre (*movement*) aux séries chronologiques existantes, l'indice *time dummy* avec période mobile (*rolling year time dummy*) est également appelé '*Time Dummy avec Movement Splice*' et est noté sous la forme $TDMS^{0,T}$. Les résultats des 4 segments sélectionnés pour un TDMS par rapport à la variante sans période mobile (TD) sont présentés ci-dessous. Pour les ordinateurs portables et les tablettes, on observe une légère différence entre la variante '*pooled*' et le '*movement splice TD index*'. Cela s'explique principalement par le taux d'attrition plus élevé de ces segments, raison pour laquelle l'estimation des coefficients de régression utilise davantage de produits différents dans les "fenêtres" du TDMS. Il s'agit, bien entendu, d'une période relativement courte et la différence s'accentuera au fur et à mesure que la période s'allonge.

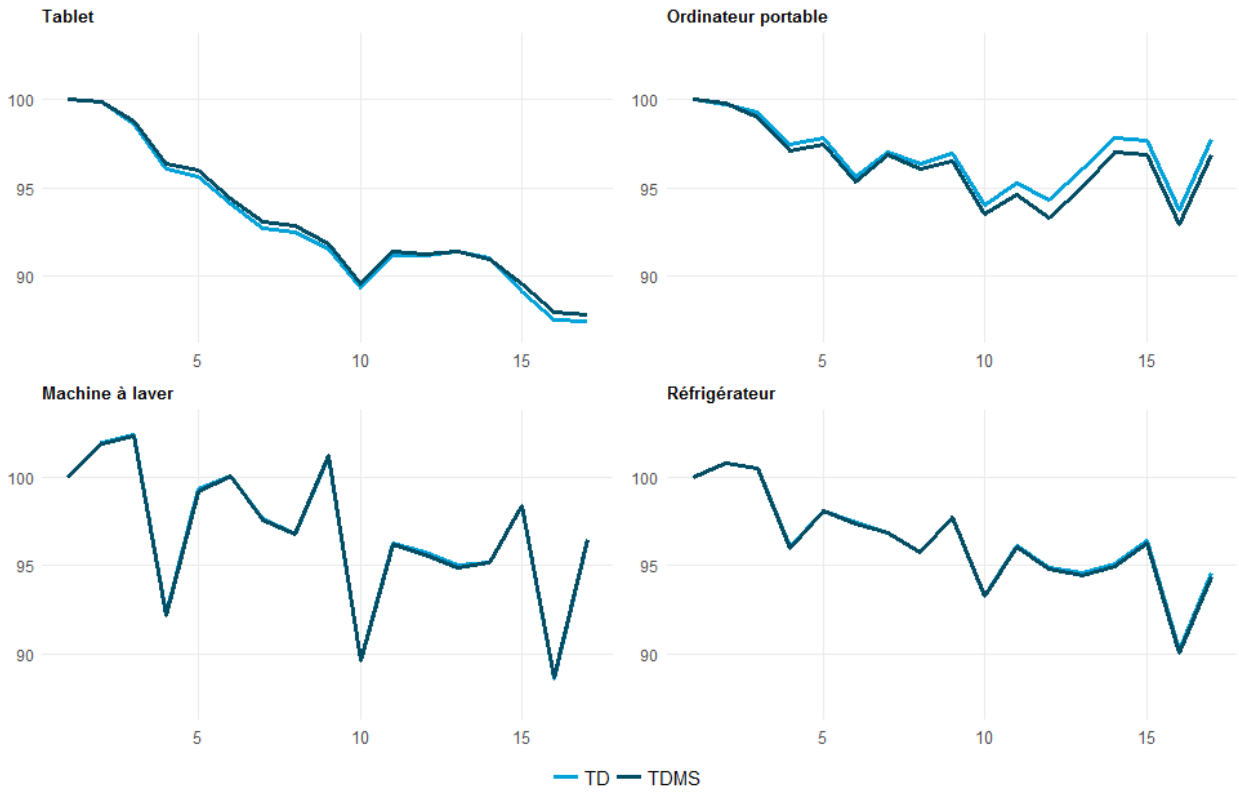


Figure 9: Comparaison entre la méthode *Pooled Time Dummy* et le *Time Dummy* avec *Movement Splice*

Si aucune information sur les caractéristiques n'est disponible, on ne peut appliquer une méthode hédonique. Les effets hédoniques $(\sum_{k=1}^{K}\hat{\beta}_k x_{ik})$ manquent alors dans l'équation standard pour un indice *time dummy* $\left(\hat{\alpha} + \sum_{t=1}^{T}\hat{\delta}^t {D_i}^t + \sum_{k=1}^{K}\hat{\beta}_k x_{ik}\right)$.

5.2.3. Fixed Effects avec Window Splice

Krsinich (2016)[11] indique que des corrections de qualité peuvent néanmoins être effectuées sur la base d'effets spécifiques au produit, les *fixed effects*' γ_i, qui remplacent les effets hédoniques manquants. De ce fait, la forme fonctionnelle d'un modèle à effets fixes (*fixed effects*) sera alors la suivante pour chaque produit i:

$$ln(p_i^{0,T}) = \hat{\alpha} + \sum_{t=1}^{T}\hat{\delta}^t {D_i}^t + \sum_{i=1}^{N-1}\hat{\gamma}_i D_i \qquad \text{Équ. 30}$$

où ${D_i}^t$ est la variable muette temporelle ('time-dummy') qui prend la valeur 1 si le produit i est disponible pendant la période t et 0 s'il n'est pas disponible, D_i est une variable muette du produit 'product-dummy' qui prend la valeur 1 si l'observation est liée au produit i. L'antilogarithme (l'exponentielle) de $\hat{\delta}^t$ donne l'indice à effets fixes (*fixed effects index*) ($FE^{0,t}$) pour la période t. Sur la base d'un raisonnement analogue développé dans les équations 10 à 24, un modèle à effet fixe (*fixed-effects model*) peut également être écrit comme suit:

$$FE^{0,T} = exp(\hat{\delta}^T) = \frac{\prod_{i \in G_T}(p_i^T)^{1/N_T}}{\prod_{i \in G_0}(p_i^0)^{1/N_0}}\, exp(\bar{\gamma}^0 - \bar{\gamma}^T) \qquad \text{Équ. 31}$$

Où $\bar{\gamma}^0 = \sum_{i \in G_0}\hat{\gamma}_i / N^0$ et $\bar{\gamma}^T = \sum_{i \in G_T}\hat{\gamma}_i / N^T$ sont les moyennes des effets fixes estimés.

Il convient également de noter qu'un article doit être disponible pendant au moins deux périodes pour avoir un impact pertinent sur un indice à effets fixes (*fixed effects index*) (par opposition à un indice hédonique *time dummy* où il peut avoir un effet immédiat). Les raisons en sont que les effets fixes sont calculés sur la base d'une régression à effets fixes, qui utilise

[11] Krsinich, F. (2016). *The FEWS Index: Fixed Effects with a Window Splice*. Journal of Official Statistics, Vol.32, No. 2, 2016, pp. 375-404.

des données de panel (observations multiples d'une même variable sur plusieurs périodes de temps) si bien que seuls les produits disponibles pendant au moins deux périodes ont un impact sur la régression (et logiquement les indices à effets fixes ultérieurs)[12].

A partir de la deuxième période où le produit est disponible, l'effet fixe de ce produit aura bien un impact sur l'indice à effets fixes qui en résultera. Plus les observations (= périodes) d'un produit sont nombreuses, plus l'effet fixe du produit converge vers sa valeur réelle. Cela implique toutefois que la méthode de travail consistant à travailler avec une année mobile (*rolling year*) telle que décrite ci-dessus n'est pas appropriée car seule l'évolution la plus récente d'une période à l'autre est couplée la série d'indices déjà existante, de sorte que la valeur actualisée de l'estimation des effets fixes du produit en question n'est pas incluse dans les indices antérieurs.

Pour résoudre ce problème, Krsinich propose d'utiliser une méthode de *window splicing*, qui s'appelle la méthode FEWS (*fixed effects with window splicing*), selon laquelle la nouvelle estimation sur 13 mois est entièrement couplée (dans le cas où T est chaque fois 13 mois) à l'indice de 12 mois auparavant, lui-même obtenu à partir de l'estimation des 13 mois antérieur. De cette façon, les valeurs actualisées des effets fixes sont bien intégrées dans l'index. L'indice FEWS pour les 13 premiers mois est calculé par défaut comme pour l'indice *time dummy*

$$ln(p_i^{0,12}) = \hat{\alpha} + \sum_{t=1}^{12} \hat{\delta}^t D_i{}^t + \sum_{i=1}^{N-1} \hat{\gamma}_i D_i \qquad \text{Équ. 32}$$

Pour le 14e mois, l'indice ($FEWS^{0,13}$) est calculé en couplant l'évolution des prix de la nouvelle période d'estimation de 13 mois à l'indice des prix de 12 mois auparavant:

$$FEWS^{0,13} = FEWS_{0,12}^{0,12} * \frac{FEWS_{1,13}^{1,13}}{100} \qquad \text{Équ. 33}$$

Dans le graphique suivant, l'indice FEWS est comparé aux indices RYGEKS et TDMS. Il est à noter que les résultats de la méthode FEWS diffèrent très peu de ceux de la RYGEKS et que les effets fixes ne peuvent quand même pas remplacer les effets hédoniques. Les méthodes hédoniques sans caractéristiques semblent donc difficilement réalisables, ce qui est conforme à la publication de de Haan et Hendriks (2013)[13].

[12] Si l'indice est calculé sur deux périodes seulement, les effets moyens estimés pendant les deux périodes $\bar{\gamma}^0$ et $\bar{\gamma}^1$ sont identiques, de sorte que la différence entre ces deux moyennes est nulle et l'indice à effets fixes (*fixed effects index*) est identique à un *matched Jevons index*.

[13] de Haan, J., Hendriks, R. (2013), *Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes*, Economic Measurement Group, Sydney, Australie.

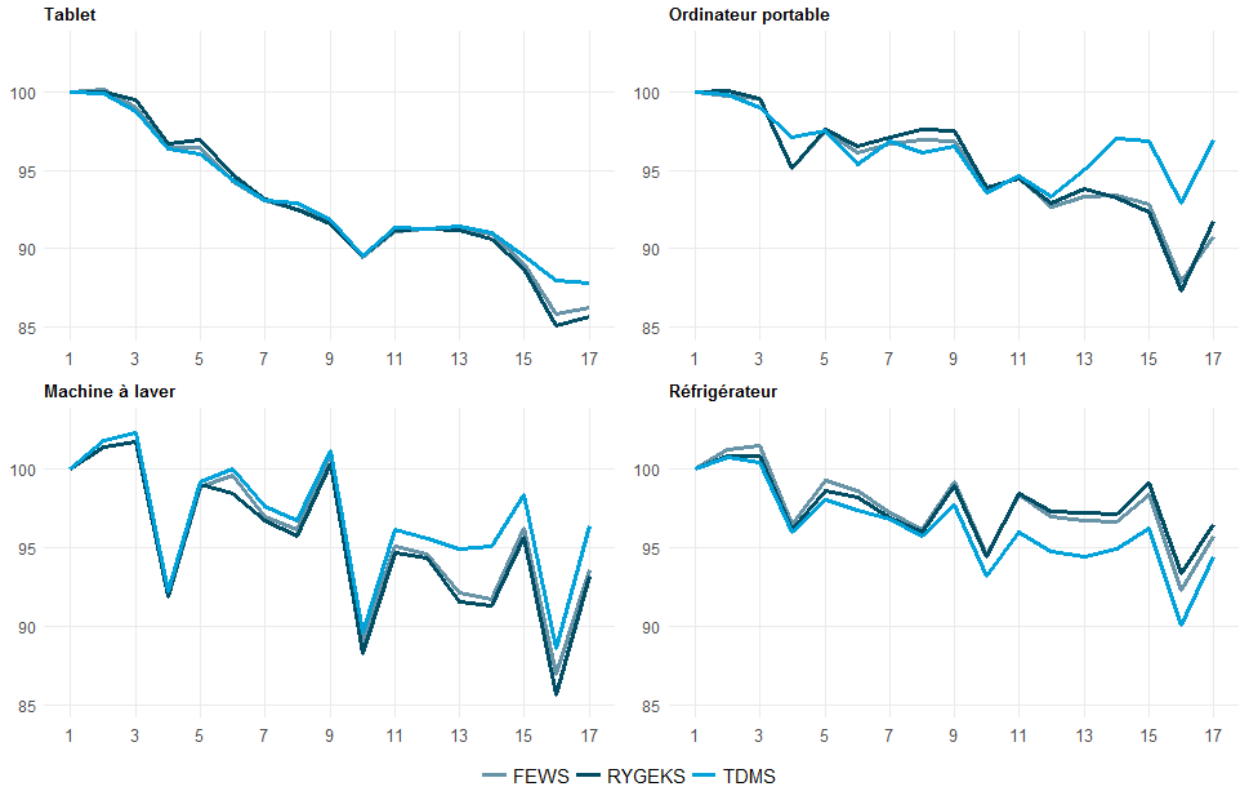


Figure 10: Comparaison entre les méthodes FEWS et RYGEKS et le *Time Dummy* avec *Movement Splice*

5.2.4. Time Dummy avec Window Splice

La méthode *window splice* mentionnée ci-dessus peut également être appliquée à la méthode hédonique *time dummy*. Tout comme la *Time Dummy with Movement Splice* (TDMS), la TDWS utilise une période mobile (*rolling window*) et peut également être généralement notée comme $TDWS^{0,T}$. Toutefois, la méthode de couplage des périodes d'estimations successives diffère.

Il est à noter que l'indice pour les 13 premiers mois est identique dans les deux cas:

$$\ln(p_i^{0,12}) = \hat{\alpha} + \sum_{t=1}^{12} \hat{\delta}^t {D_i}^t + \sum_{k=1}^{K} \hat{\beta}_k x_{ik} \qquad \text{Équ. 34}$$

L'estimation pour les 13 prochains mois est également identique.

$$ln(p_i^{1,13}) = \hat{\alpha} + \sum_{t=1}^{12} \hat{\delta}^t {D_i}^t + \sum_{k=1}^{K} \hat{\beta}_k x_{ik} \qquad \text{Équ. 35}$$

La différence entre les deux méthodes à période mobile (*rolling window*) réside dans la façon dont l'indice du 14e mois est calculé. Pour rappel, dans le *movement splice*, seule l'évolution mensuelle des deux derniers mois de la nouvelle période d'estimation est couplée à l'ancien indice (cf. équation 28). Alors que, dans la méthode *window splice*, la nouvelle estimation sur 13 mois est entièrement couplée (dans le cas où T est chaque fois 13 mois) à l'indice de 12 mois plus tôt qui a été obtenu à partir de l'estimation des 13 mois précédents:

$$FEWS^{0,13} = TDWS_{0,12}^{0,12} * \frac{TDWS_{1,13}^{1,13}}{100} \qquad \text{Équ. 36}$$

Cela signifie que les indices déjà publiés restent inchangés dans les deux cas et sont même identiques avec les deux méthodes pour les 13 premiers mois. A partir du 14e mois, l'indice est normalement différent parce que la méthode *window splice* pour la période T +1 (=0,..., T,T+1) prend en compte l'évolution du prix basée sur de nouveaux paramètres entre le mois 1 et T+1,

alors que, dans la méthode *movement splice*, seule l'évolution du prix basée sur les nouveaux paramètres entre T et T+1 est prise en compte.

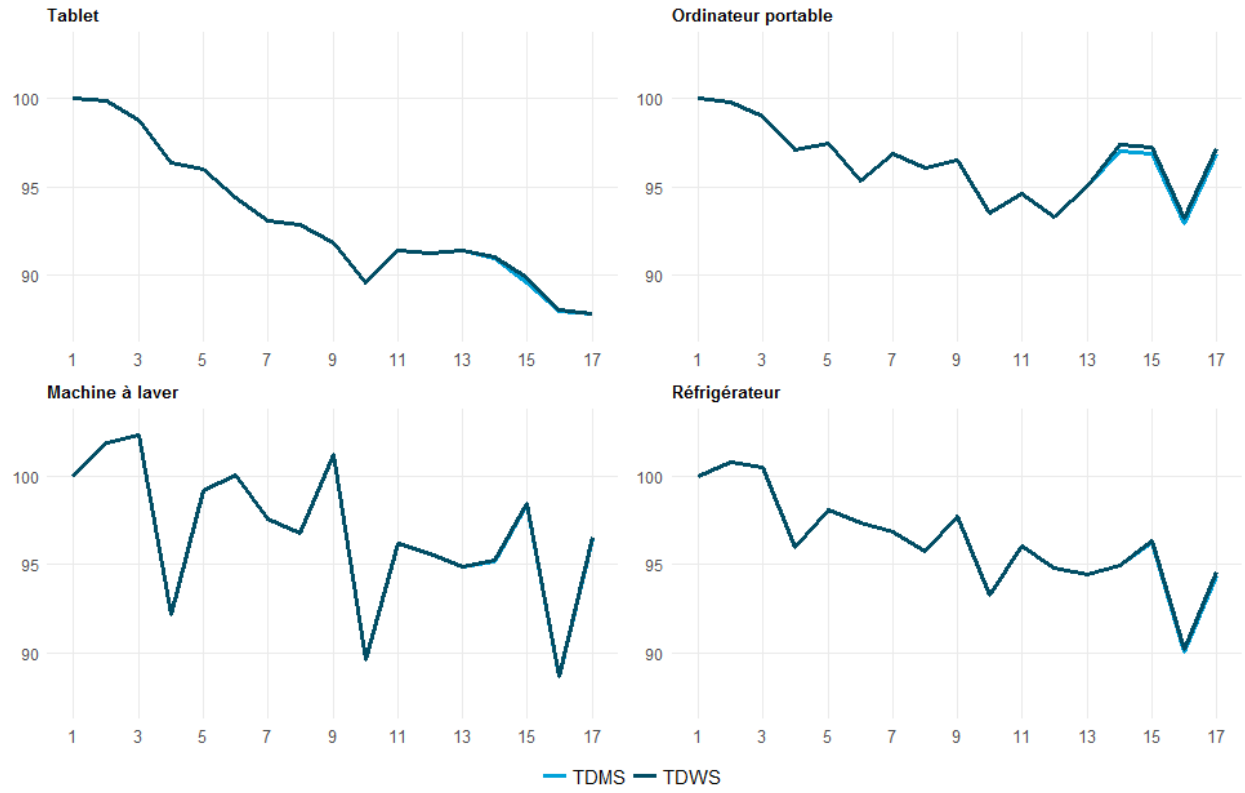


Figure 11: Comparaison entre *Time Dummy* avec *Movement Splice* et *Time Dummy* avec *Window Splice*

La figure ci-dessus montre une petite différence entre la méthode *time dummy* avec *movement splice* et la méthode avec *window splice*. Comme l'évolution des prix au cours des 13 premiers mois est identique, nous ne constatons une différence qu'à partir du 14^{e} mois. Si les données sont recueillies et traitées sur une période plus longue (24 mois), la différence est un peu plus nette. On constate également que les indices TD du réfrigérateur sont aussi supérieurs à ceux du RYGEKS:

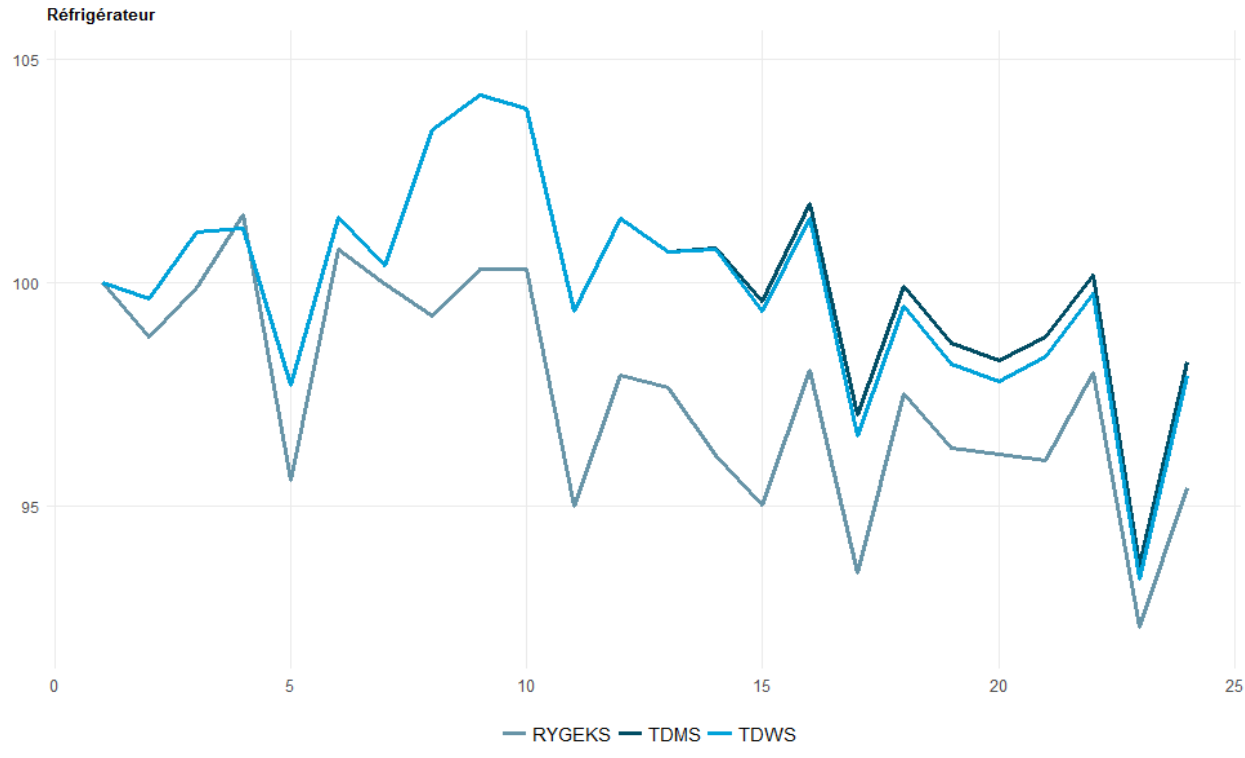


Figure 12: Comparaison RYGEKS, TDMS et TDWS pour les réfrigérateurs sur une période de 24 mois

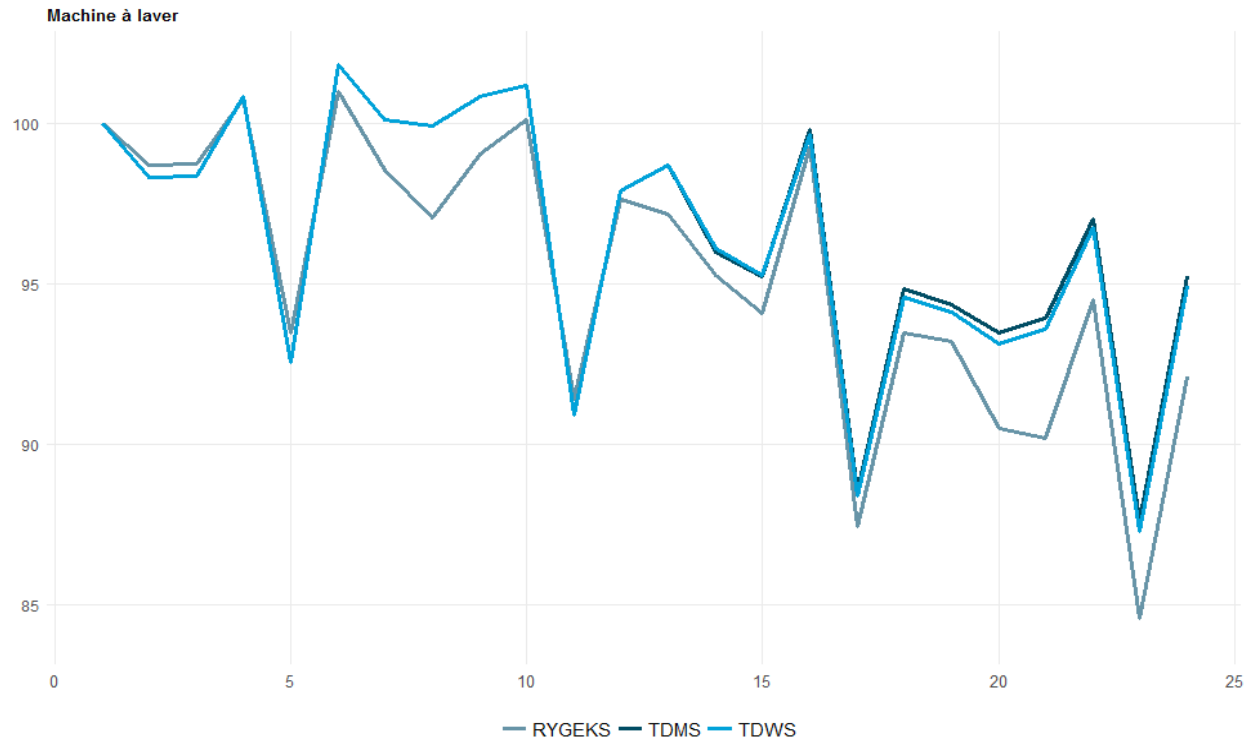


Figure 13: Comparaison RYGEKS, TDMS et TDWS pour les machines à laver sur une période de 24 mois

L'inconvénient de la méthode est qu'il est plus difficile d'interpréter l'évolution mensuelle entre T en T+1 car elle est calculée sur la base de deux périodes d'estimation comprises entre 0 à Tet 1 à T+1 sur la base de deux modèles et ne reflète donc pas l'évolution mensuelle des prix d'un seul et même modèle comme dans le *movement splice*. Toutefois, la méthode *window splice* présente l'avantage de prendre en compte l'impact des paramètres modifiés sur toute la période. Un autre avantage est qu'en cas d'intégration d'une nouvelle caractéristique dans le modèle de régression, l'effet de cet ajout sur toute la période est pris en compte et pas seulement son effet sur les deux derniers mois comme c'est le cas avec le *movement splice*. Par conséquent, la méthode *Time Dummy* avec *Window Splice* est actuellement privilégiée.

5.3. Réserves concernant les méthodes hédoniques

Les méthodes hédoniques ne sont actuellement appliquées presque nulle part en Europe. Dans la zone euro, elles sont actuellement utilisées uniquement pour les voitures d'occasion, les tablettes et les ordinateurs en Allemagne, pour les livres, les machines à laver et les réfrigérateurs en France et uniquement pour les voitures d'occasion aux Pays-Bas. L'un des principaux problèmes liés à l'application des méthodes hédoniques est qu'elles exigent un assez grand nombre de données et qu'elles nécessitent donc beaucoup de travail par rapport à l'importance de ces produits dans le panier de l'indice.

Par exemple, il faut noter non seulement les caractéristiques des produits mais l'échantillon doit également être suffisamment grand pour garantir des degrés de liberté suffisants pour obtenir des estimations fiables des paramètres dans l'équation de régression. Dès lors, la collecte des données est souvent sous-traitée pour ces groupes de produits (par exemple en Allemagne et aux Pays-Bas), voire même l'ensemble du calcul de l'indice dans le cas des voitures d'occasion aux Pays-Bas.

Grâce au *webscraping*, il devient un peu plus aisé d'obtenir suffisamment d'observations et d'enregistrer les caractéristiques. Le traitement des données et la rédaction des scripts pour le *scraping* des sites internet nécessitent bien sûr aussi beaucoup de travail, raison pour laquelle cette méthode n'est appliquée ici aussi qu'à certains groupes de produits.

Il convient également de noter que les variables ne sont souvent pas répartie de manière continue (seulement quelques valeurs par variable, et elles ne sont certainement pas le résultat d'un processus stochastique gaussien). Pensez, par exemple, à la consommation d'eau ou d'électricité d'une machine à laver, où, à un moment donné, elle a une répartition discrète avec de faibles variations. Une plus grande variance par variable ne s'obtient que sur une plus longue période, l'équation de régression initiale devenant moins représentative. Dans ce cas, cependant, le *webscraping* présente un avantage par rapport aux méthodes hédoniques utilisées pour la collecte de données classique, dans lesquelles les comparaisons ne sont souvent estimées que sporadiquement en raison du processus à forte intensité de main-d'œuvre de collecte d'un nombre suffisant

de prix et de caractéristiques. Avec le *webscraping*, il est possible d'appliquer une équation de régression qui est en permanence à jour. Grâce au *scraping* régulier des sites internet, les données deviennent de plus en plus nombreuses et précises. Entre-temps, les scripts ont été complétés par de nouveaux segments tels que les home cinémas, les télévisions et les smartphones.

6. ÉTUDE DE CAS – VOITURES D'OCCASION

Les voitures d'occasion constituent un segment qui n'est actuellement pas repris dans les indices des prix à la consommation national et harmonisé, bien que son poids dépasse le seuil de 1 pour mille, comme décrit dans la réglementation sur l'IPCH.

Cette réglementation exige que tous les biens de consommation et services pour lesquels les dépenses des ménages dépassent ce seuil soient repris dans le calcul de l'IPCH. Jusqu'à présent, Statbel n'inclut pas ce segment de consommation, car il est difficile d'obtenir des données permettant de calculer correctement un indice pour les voitures d'occasion.

Il n'est pas non plus possible pour ce segment de calculer un indice qui, d'une manière ou d'une autre, ne tient pas compte des différences entre les voitures d'occasion offertes, étant donné que naturellement, l'âge, le nombre de kilomètres, le type de carburant, etc. d'un certain type de voiture diffèrent à chaque fois.

Le webscraping permet de calculer plus facilement cet indice: suffisamment de prix (et de caractéristiques) sont disponibles sur des sites spécialisés dans l'offre de voitures d'occasion. Un échantillon reprenant les voitures d'occasion les plus populaires a été établi à partir de la base de données des immatriculations de voitures du Service public fédéral Mobilité. Des données sont ensuite extraites pratiquement quotidiennement par *scraping* pour cet échantillon. Les prix d'une même observation apparue plusieurs fois sont ramenés à un seul prix.

Après le nettoyage des données (suppression des valeurs aberrantes pour les prix l'âge des voitures, le nombre de kilomètres, etc.), l'indice est estimé sur la base de l'indice hédonique *time dummy* décrit ci-dessus. Ici aussi, la variante semi-logarithmique est utilisée. Les données des différents mois sont utilisées dans l'équation de régression, dans laquelle chaque mois (= période) se voit attribuer une variable indicatrice (*dummy*) (il s'agit donc ici de la variante *pooled* de la régression).

L'indice obtenu est illustré dans le graphique suivant, qui montre que l'évolution des prix mesurée est assez stable. Ce n'est pas entièrement illogique, car il serait surprenant que le prix varie fortement de mois en mois pour les voitures "standardisées".

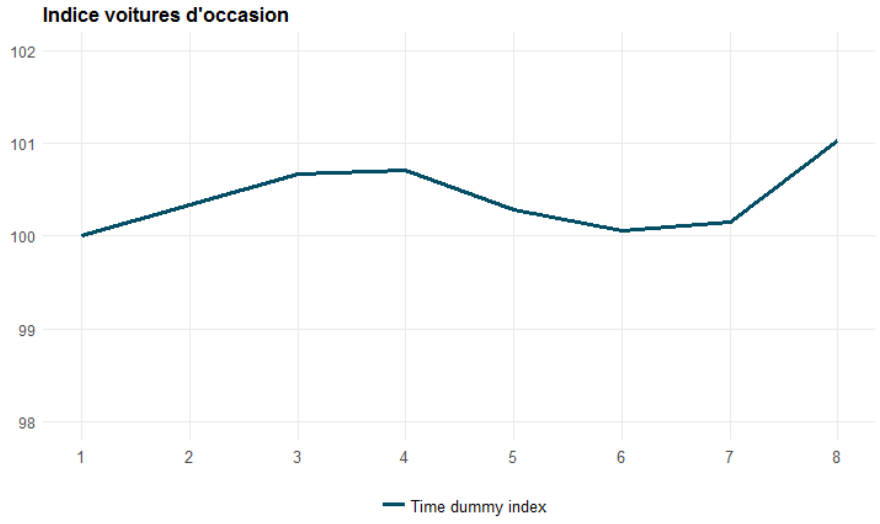


Figure 14: Indice *time dummy* pour les voitures d'occasion

Il est conseillé d'examiner la manière dont l'indice *time dummy* corrige les différences de qualité entre les voitures. Il ressort du chapitre précédent que ce type d'indice peut être divisé pour une certaine période en un indice brut (rapport des moyennes géométriques des prix), qui ne tient pas compte des différences de qualité des observations entre les périodes, et un facteur qui effectue un ajustement en ce qui les différences de qualité entre les périodes sur la base des modifications des caractéristiques. La formule standard pour la période t est:

$$TD^{0,t} = exp(\hat{\delta}) = \frac{\prod_{i \in G_t} (p_i^1)^{1/N_1}}{\prod_{i \in G_0} (p_i^0)^{1/N_0}} \, exp\left(\sum_{k=1}^{K} \hat{\beta}_k (\bar{x}_k^0 - \bar{x}_k^t)\right) \qquad \text{Équ. 37}$$

Cette équation peut être réécrite comme suit, en prenant simplement l'inverse du deuxième composant de l'équation ci-dessus pour obtenir un ratio au lieu d'une multiplication:

$$TD^{0,t} = \frac{\prod_{i \in G_t} (p_i^1)^{1/N_1}}{\prod_{i \in G_0} (p_i^0)^{1/N_0}} \Bigg/ \exp\left(\sum_{k=1}^{K} \hat{\beta}_k (\bar{x}_k^0 - \bar{x}_k^t)\right)^{-1} = \frac{\text{Unadjusted index}}{\text{Quality factor index}}$$

Équ. 38

Par conséquent, $TD^{0,t}$ est donc égal au ratio entre les deux indices, à savoir un indice de Jevons (*unadjusted index*) et un indice de la qualité des voitures dans l'échantillon (*quality factor index*). De cette manière, on obtient un type d'indicateur "nominal" corrigé par un déflateur. La répartition entre les deux indices est illustrée dans le graphique suivant.

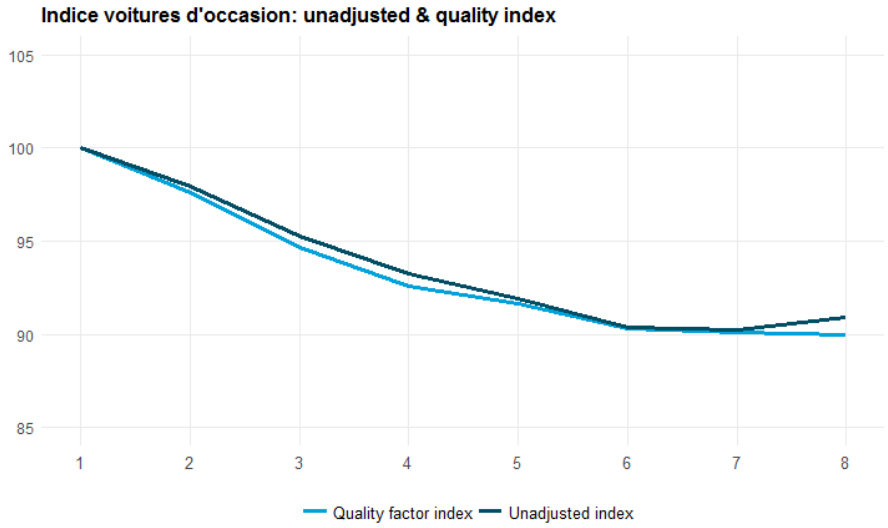


Figure 15: *Unadjusted* et *Quality factor index* pour les voitures d'occasion

Les deux indices enregistrent une baisse, ce qui ne veut pas seulement dire que les prix moyens de l'échantillon de voitures offertes ont diminué, mais aussi que la qualité de ces voitures a diminué. La division des deux indices est égale à l'indice *time dummy*. Par exemple, on constate dans la période 8 que l'indice de qualité (*quality factor index*) a diminué de manière plus significative que l'indice brut (*unadjusted index*) des moyennes géométriques des prix. Par conséquent, l'indice *time dummy* s'inscrit à la hausse pour atteindre un niveau de 101.

Compte tenu des résultats obtenus et de la réglementation sur l'IPCH, ce segment de consommation sera normalement inclus à partir de 2019. Logiquement, un choix doit donc être fait en ce qui concerne la méthode du *window splicing* ou du *movement splicing* (ou une autre méthode de *splicing* ne relevant pas du champ d'application de ce texte). Pour l'indice des prix à la consommation national, cela dépend de l'avis de la Commission de l'indice.

7. ÉTUDE DE CAS – LES HÔTELS

Depuis quelques mois, les données relatives aux chambres d'hôtel sont collectées par scraping pour trois destinations belges: un weekend à la côte belge, un weekend dans les Ardennes et un weekend à Bruxelles.

Dans le cadre des relevés de prix manuels, un échantillon d'hôtels est constitué et les prix sont collectés une fois par mois pour une réservation 4 semaines avant la date d'arrivée (1 relevé de prix par hôtel). La réservation virtuelle est faite pour une chambre de deux personnes et pour un séjour de deux nuits. En principe, le type de chambre et les autres options restent stables. Naturellement, cela n'est pas toujours réalisable d'un point de vue technique, étant donné que cela dépend de la disponibilité des chambres. L'échantillon est composé de hôtels pour Bruxelles, pour la côte belge et pour les Ardennes.

Lors du webscraping, les données sont collectées tous les jours pour un séjour réservé (virtuellement) 4 et 8 semaines avant l'arrivée. Cela donne 1 prix par hôtel par date de réservation pour un séjour 4 et 8 semaines après celle-ci. Les réservations virtuelles sont ici toujours effectuées pour une arrivée le vendredi et un départ le dimanche, et comprennent également le petit-déjeuner et l'annulation gratuite. Par ailleurs, une stratification supplémentaire par région (côte et Ardennes) est réalisée. Dès lors, 5 villes côtières et 9 destinations ardennaises se trouvent dans l'échantillon. Les hôtels bruxellois sont limités au centre-ville. De plus, une stratification supplémentaire est encore réalisée au niveau de la notation en étoiles: 2, 3 et 4 étoiles. La stratification complète peut être illustrée comme suit:

Région

⬇

Destination

⬇

Nombre de semaines entre la réservation et le séjour

⬇

Nombre d'étoiles

Il convient de préciser que, conformément à la réglementation de l'IPCH relative au calcul des indices des services, l'évolution des prix mesurée est incluse dans l'indice du mois au cours duquel le service commence effectivement. Cela signifie donc que, pour un séjour dans un hôtel, le jour d'arrivée à l'hôtel conformément à la réservation déterminera le mois durant lequel le prix sera inclus dans l'indice. Le raisonnement derrière est que le mois du séjour est plus déterminant pour l'évolution des prix que le moment de la réservation. En d'autres termes, l'évolution des prix mesurée entre les mois dépend plus du mois de séjour que du nombre de semaines entre la réservation et le séjour. Concrètement, le prix d'une réservation pour un séjour qui débute en avril, qu'elle ait été faite 4 ou 8 semaines à l'avance, est inclus dans l'indice d'avril. Ce dernier contiendra donc les prix des réservations qui ont été effectuées en février, mars ou même avril. De même, la date de départ n'est pas pertinente. Un séjour qui commence le 30 avril et se termine le 2 mai sera inclus dans l'indice d'avril.

Un indice de Jevons est utilisé au niveau le plus bas du modèle de stratification schématique visé ci-dessus. La moyenne géométrique des prix au niveau le plus bas est ensuite comparée au prix tel qu'enregistré au cours du mois de base (premier mois). Les indices résultants sont alors progressivement agrégés à un niveau supérieur afin d'obtenir de cette manière un indice par région. Grâce au *webscraping*, on ne calcule donc plus un indice par hôtel, mais par strate.

Les strates ainsi que la limitation de la réservation à un certain type de chambre avec les mêmes options, de même que la période du séjour, permettent d'obtenir un service homogène au niveau le plus bas ne nécessitant aucune adaptation de qualité. Grâce à cette méthode, un prix par strate est également toujours disponible, de même qu'un "prix de base" par strate qui peut être calculé en décembre et auquel peuvent être comparés les prix de l'année en cours. Par conséquent, l'inclusion des prix des hôtels pour lesquels aucune chambre n'est disponible pendant une certaine période ou des prix des nouveaux hôtels ne pose aucun problème méthodologique. Dans ce cas, il n'est pas non plus nécessaire de travailler avec une année mobile (*rolling year*).

Le *webscraping* permet également d'augmenter facilement le nombre d'hôtels dans l'échantillon. De même, le *scraping* quotidien permet bien entendu d'obtenir un nombre d'observations des prix bien plus élevé que la méthode manuelle. Le nombre de prix obtenus par *scraping* pour 1 mois s'élève en moyenne à:

Région	Nombre de prix
Bruxelles	2 662
Côte	12 614
Ardennes	23 552

Cela donne les résultats suivants:

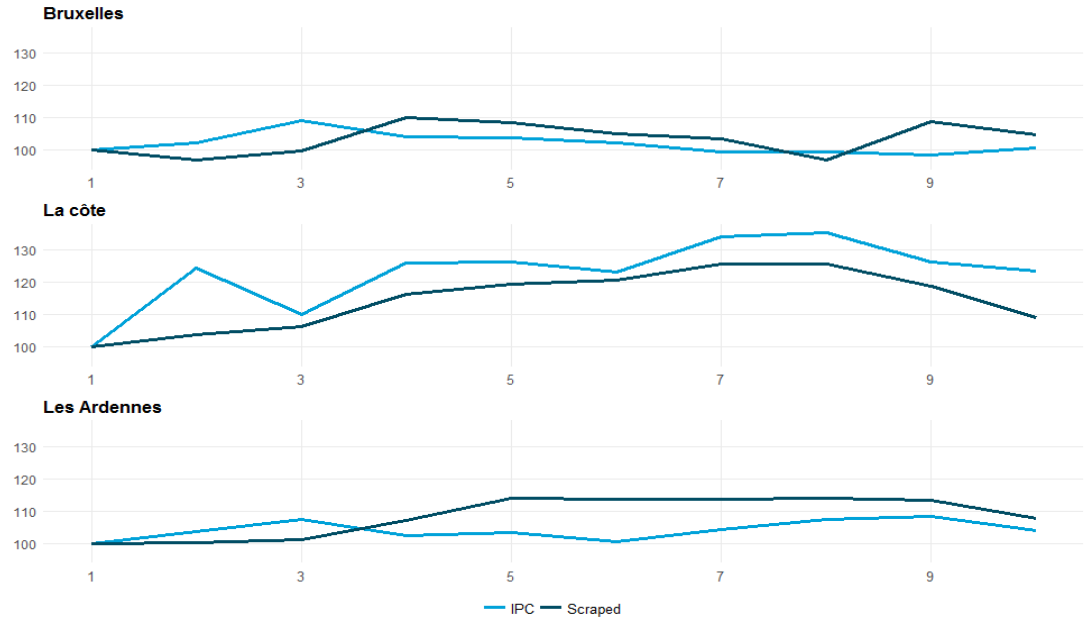


Figure 16: Comparaison de l'évolution de l'indice entre les données issues du *webscraping* et les chiffres de l'IPC

Dans l'ensemble, les données issues du *webscraping* et les chiffres officiels de l'IPC affichent une même tendance dans l'évolution de l'indice. Les résultats pour Bruxelles sont même plus logiques étant donné qu'en 2017, les vacances de Pâques tombaient entièrement en avril (mois 4), de sorte qu'une hausse de l'indice était attendue en avril. Les chiffres officiels montrent toutefois une augmentation en mars (mois 3) et une diminution en avril (mois 4).

Entre-temps, le nombre de destinations pour lesquelles des données sont collectées par *webscraping* a encore augmenté, avec l'ajout de 5 villes flamandes et de 7 villes wallonnes. Des recherches complémentaires sont prévues en 2018.

Les indices obtenus par *webscraping* donnent un bon résultat et semblent moins volatiles que les indices officiels. Cela peut s'expliquer par un échantillon plus grand, la stratification et la limitation de la réservation aux chambres avec petit-déjeuner et annulation gratuite. Étant donné que la méthode manuelle ne relève les prix qu'une seule fois par mois pour le même hôtel, il est possible qu'un certain type de chambre ne soit plus disponible au moment de la réservation. Par conséquent, une autre chambre sera choisie dans le même hôtel, ou éventuellement la même chambre avec d'autres conditions (par exemple la même chambre pour laquelle l'annulation gratuite ne serait pas comprise).

8. ÉTUDE DE CAS – LES CHAMBRES D'ÉTUDIANTS

La location de chambres d'étudiants n'intervient actuellement pas dans le calcul de l'indice. Pour le moment, le segment de la location se compose uniquement des logements privés et sociaux Étant donné que les dépenses relatives aux chambres d'étudiants semblent tout de même importantes, il est intéressant de mener une étude sur ce segment.

Contrairement aux locations privées, il n'est pas possible, pour les chambres d'étudiants, de consulter une base de données administratives des contrats de location enregistrés, étant donné que ces contrats ne sont pas souvent enregistrés et qu'il est difficile de reconnaître dans la base de données si un contrat est conclu ou non pour la location d'une chambre d'étudiants. Lors des relevés de prix des locations privées, un échantillon est tiré à partir de la base de données administratives des contrats de location, et les locataires (les sociétés de logements sociaux pour les locations sociales) sont contactés.

Les données de contact des étudiants qui louent un kot pourraient éventuellement être obtenues auprès des universités, mais on s'attend à un faible taux de réponse. En effet, l'enquête ordinaire sur les loyers privés enregistre également un taux de réponse très faible, alors que les personnes concernées reçoivent normalement plus de courrier à leur adresse.

La courte durée des baux de location des chambres d'étudiants constitue un autre problème. L'enquête sur les loyers privés fonctionne au moyen d'un système de panels, dans lequel un même bien locatif fait l'objet d'un suivi dans le temps via une enquête auprès du locataire (seule l'interrogation du locataire est reprise que dans l'A.R. correspondant). Cela peut poser problème pour les chambres d'étudiants car les contrats de location privée normaux sont en principe également caractérisés par une longue durée (3 ans avec à chaque fois une prolongation de 3 ans), alors que les chambres d'étudiants sont louées pour un an (ou même 10 mois). Afin de pouvoir recourir à la même méthodologie que pour l'enquête sur les loyers privés, le nouvel étudiant doit à chaque fois être disposé à participer à l'enquête. En raison de la courte durée et des étudiants qui abandonnent ou terminent leurs études, le nombre de chambres à louer est chaque année important, et ce pendant un nombre limité de mois (juin - septembre). Pour ces raisons, il est intéressant pour ce segment de collecter des données par *webscraping* via des sites proposant des chambres d'étudiant.

Les données suivantes relatives aux chambres d'étudiants ont été obtenues par *webscraping* pour différentes villes estudiantines: prix, taille et adresse. Un filtre est ensuite appliqué à l'adresse via géocodage afin de minimiser l'impact de la distance entre le bien locatif et le campus.

Un exemple de géocodage est présenté dans le graphique suivant. Le résultat après application du filtre de la distance jusqu'au campus se trouve à droite.

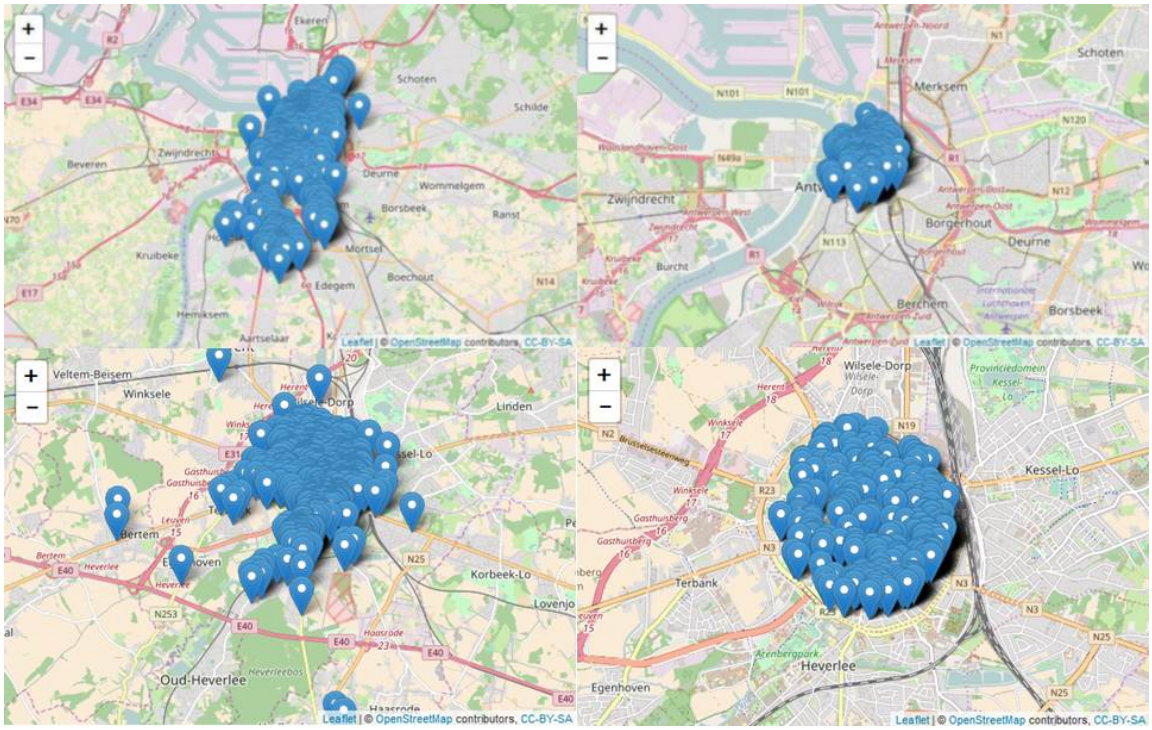


Figure 17: Exemple de géocodage pour 2 villes estudiantines (Anvers et Louvain)

L'indice a été calculé pour les villes d'Anvers et de Louvain en comparant les prix de juin 2017 à ceux de juillet 2016. Les deux villes affichent des résultats similaires:

Période	Ville 1	Ville 2
07-2016	100	100
06-2017	102,13	102,33

Étant donné que la période de collecte des données est assez courte (quelques mois seulement par an), il reste difficile de tester fréquemment le calcul de l'indice des chambres d'étudiants. Entre-temps, les scripts ont été élargis pour inclure d'autres villes estudiantines supplémentaires. Une évaluation plus approfondie est prévue cette année.

9. ÉTUDE DE CAS – LES CHAUSSURES

Un *scraping* des sites web des plus grands magasins de chaussures en Belgique a été effectué. Ces magasins possèdent une boutique physique et une boutique en ligne. Ce *scraping* a été effectué plusieurs fois par semaine. Il s'agissait ici aussi d'un *webscraping* en masse, ce qui signifie qu'un *scraping* a été effectué sur tous les produits du site web et qu'aucune sélection (ou limitation) n'a été faite à l'avance. Le nettoyage et la sélection des données ont été réalisés pendant la phase d'analyse.

Le *scraping* en masse diffère de la manière dont le *scraping* des produits électroniques grand public est effectué, pour lequel des produits spécifiques sont sélectionnés. Dans ce cas, la limitation est principalement utilisée pour pouvoir effectuer un *scraping* sur toutes les caractéristiques, et parce que ces produits sont normalisés sur différents sites. Les caractéristiques des chaussures sont toutefois limitées. Par conséquent, aucune limitation préalable n'est nécessaire. Par ailleurs, il serait également difficile de procéder à une limitation à l'avance, étant donné que l'offre de types de chaussures varie d'un site à l'autre.

Des produits tels que les chaussures (et les vêtements) enregistrent, encore plus que les produits électroniques grand public, un taux d'attrition (évolution des produits) élevé. Cela signifie que des produits apparaissent et disparaissent souvent de la gamme (par exemple en raison de la modification de certaines tendances de mode ou à la suite d'un changement de saison). Le graphique suivant montre le nombre d'articles pouvant être reliés à la période 1 pour une chaîne. Après 4 mois, le nombre d'articles correspondant tombe à moins de la moitié. Il diminue ensuite pour atteindre 9 % au mois 9, pour à nouveau augmenter au cours des mois 10 à 12 (un an plus tard, donc plus ou moins à la même saison), après quoi il diminue à nouveau à 6 %.

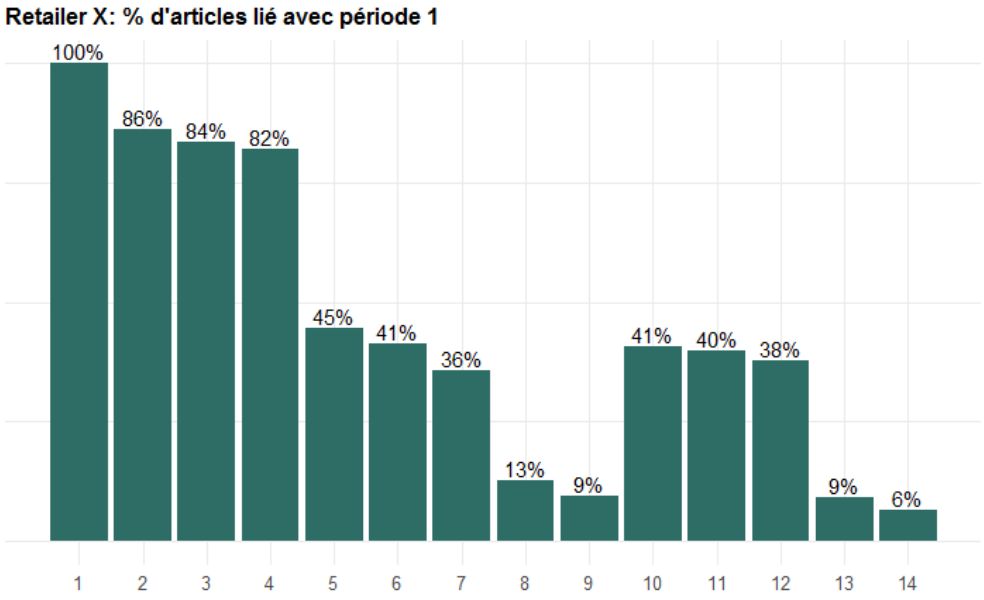


Figure 18: Taux d'attrition des chaussures

La dynamique des données issues du *scraping* peut également être visualisée. Le schéma suivant montre la disponibilité d'un produit. Les articles se trouvent sur l'axe vertical et le temps sur l'axe horizontal. Chaque pixel gris indique la disponibilité d'un produit sur le site web. Différents groupes peuvent être observés (modifications de la gamme de produits, cycles saisonniers, etc.). Le nombre de produits disponibles tout au long de la période, caractérisé par une ligne grise, est très limité.

Figure 19: Visualisation de la disponibilité des chaussures

L'évolution totale des produits est illustrée dans le graphique suivant pour deux détaillants. Le nombre total de produits est donné par rapport au nombre maximum de produits disponibles pendant toute la période (100). De même, le nombre de produits nouveaux et disparus par période est donné par rapport au nombre maximal de produits disponibles. Nous pouvons donc conclure que le nombre total de produits offerts sur le site du détaillant X varie fortement en fonction de la période, alors que ce nombre est plus stable pour le détaillant Y.

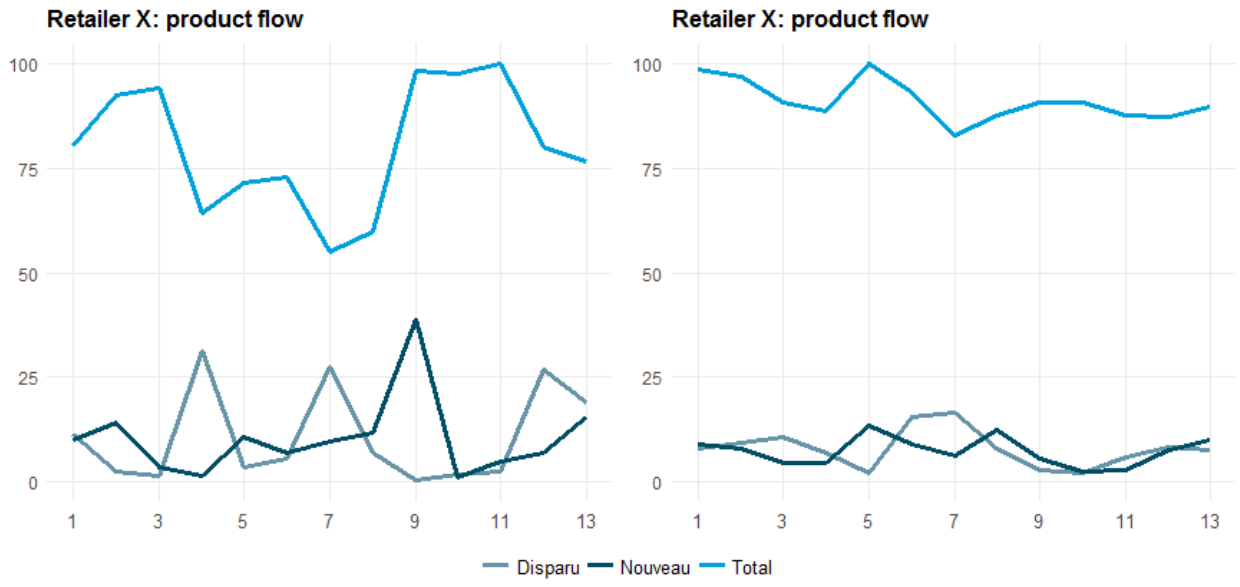


Figure 20: Nombre de produits (nombre total, produits nouveaux et disparus) chez deux détaillants

Outre ce caractère dynamique, le secteur présente une autre caractéristique: lorsque les produits disparaissent de la gamme, ils affichent généralement un prix largement inférieur à celui auquel ils ont été mis sur le marché.

Si le calcul de l'indice était basé sur l'approche du *matched model*, selon laquelle les articles disponibles lors de périodes consécutives sont couplés, on obtiendrait un indice caractérisé par une dérive négative, comme le montre le graphique ci-dessous pour le détaillant X. La dérive négative est encore plus prononcée que pour les produits électroniques grand public. Tant les chaussures pour hommes que celles pour femmes ont enregistré une baisse au cours des périodes 6 et 7 en raison des soldes. Bon nombre de ces produits disparaissent de la gamme après ces soldes (voir le graphique du flux de produits ci-dessus). Par conséquent l'indice ne revient pas à un niveau plus élevé tel que celui de la période 3 par exemple.

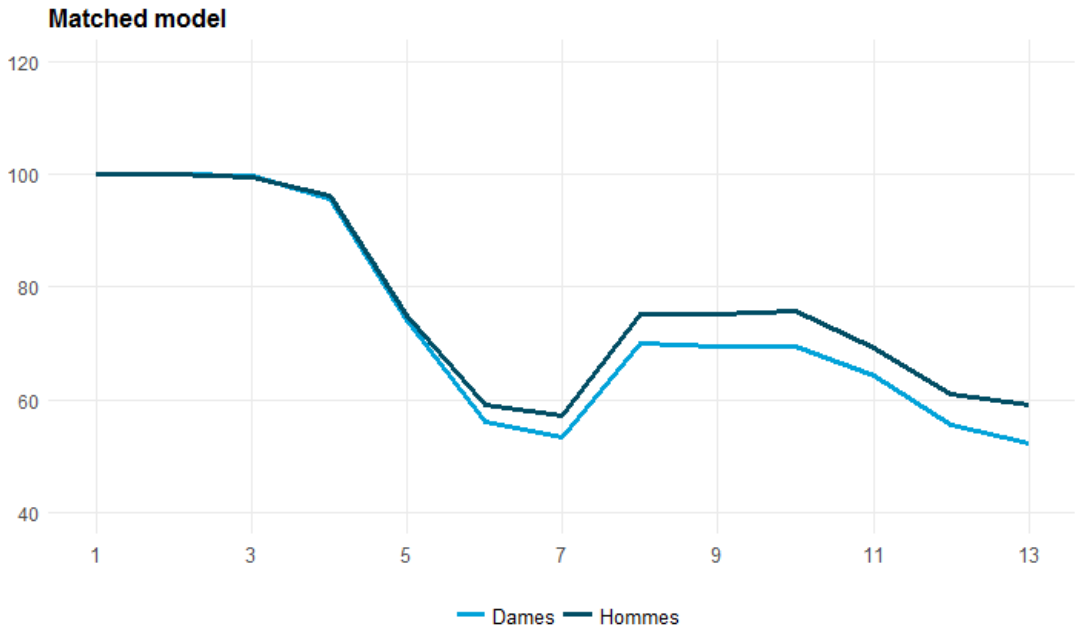


Figure 21: Dérive négative du *matched model*

Ce problème, à savoir que des articles quittent la gamme de produits à des prix très réduits et n'y reviennent ensuite plus, ne peut ici non plus être résolu en travaillant avec une méthode multilatérale telle qu'un indice GEKS (dans ce cas un indice GEKS-J) ou un indice FEWS. Ces deux indices illustrés ci-dessous évoluent de manière similaire au *matched model Jevons index*.

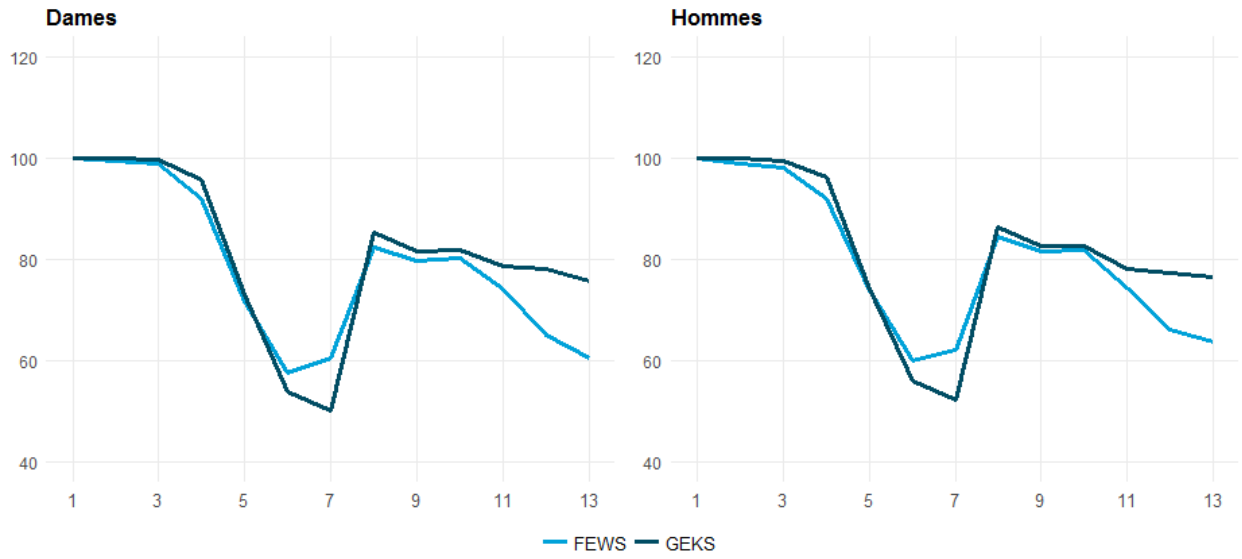


Figure 22: Dérive négative des indices FEWS et GEKS

Pour éviter la dérive négative, on applique l'approche du *non-matched model*, étant donné que les caractéristiques disponibles ne sont pas suffisantes pour permettre une correction hédonique pertinente. Pour chaque site, une stratification est d'abord réalisée pour les chaussures d'hommes et de femmes. Ensuite, chaque segment est divisé par type de chaussure (et ce à l'aide de la classification disponible sur le site web du détaillant, étant donné que l'offre peut varier d'un site à l'autre). Un indice de Jevons simple a été calculé à ce niveau le plus bas pour tous les articles. La différence entre la manière dont

l'indice est calculé pour les chaussures d'hommes et de femmes (faisant toutes deux partie d'un groupe de la COICOP à 5 positions) via *webscraping* et les relevés classiques en magasin (indice officiel actuel) est illustrée dans le schéma suivant:

Relevés de prix classiques (indice officiel)	Indice du *webscraping*
COICOP à 5 positions	COICOP à 5 positions
⬇	⬇
Échantillon des segments de consommation	Détaillant
⬇	⬇
Indice de Jevons	Segments de consommation non harmonisés
	⬇
	Indice de Jevons

L'approche classique consiste en une approche plus ascendante (*bottom-up*), dans laquelle les prix sont relevés pour un certain type de chaussure au moyen d'une définition et l'indice résultant est calculé sur la base de l'indice de Jevons. Le *webscraping* applique une approche plus descendante (*top-down*), dans laquelle on part de toutes les chaussures disponibles qui sont ensuite agrégées par détaillant pour obtenir un indice pour les chaussures d'hommes et de femmes. Un indice global peut alors être calculé.

Si l'on compare l'indice obtenu par *webscraping* aux indices officiels, on constate une tendance similaire, bien qu'il y ait une différence de niveau de l'indice pendant certains mois.

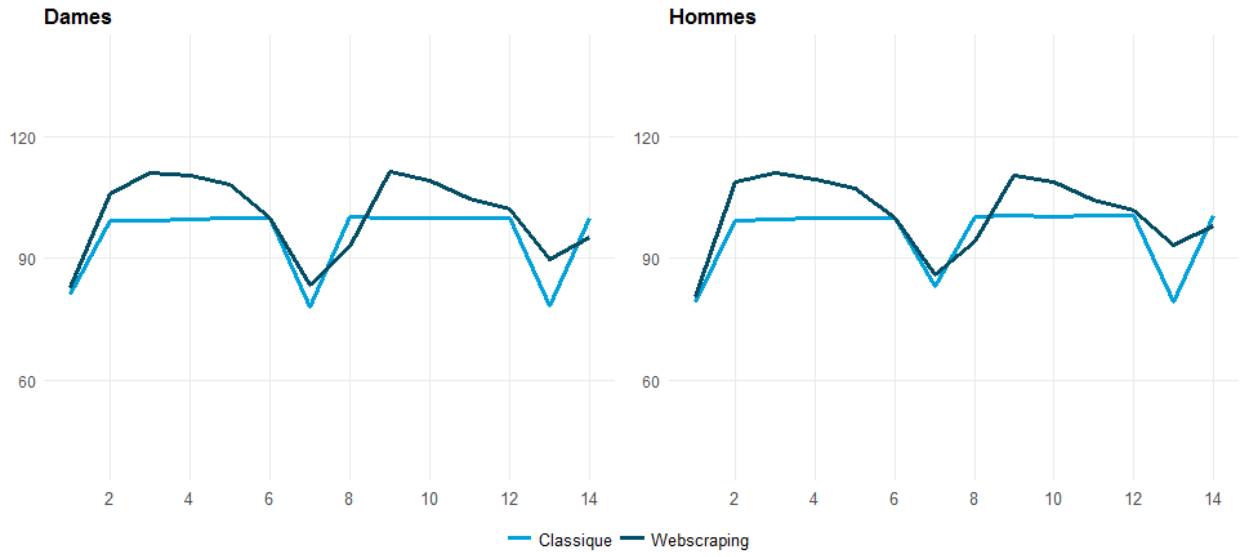


Figure 23: Comparaison des indices entre la méthode classique et le *webscraping*

Cette différence peut s'expliquer par le fait que les relevés de prix classiques n'incluent les soldes qu'en janvier et juillet.

Cette manière de faire est appliquée parce que ces soldes reviennent chaque année et durent un mois entier, ce qui rend la planification des relevés de prix plus facile. Aucun prix n'est relevé en ce qui concerne les réductions très temporaires qui ne sont pas accessibles à tous. Les données du *webscraping*, quant à elles, tiennent compte de ces réductions. Le détaillant X propose des réductions au mois 6 (qui est également le mois durant lequel l'indice est fixé à 100 parce qu'il s'agit aussi du mois de base pour les prix classiques), ce qui relève le niveau de l'indice des autres mois parce qu'il est comparé à ce faible niveau de prix. Si nous éliminons cette réduction du détaillant X, nous pouvons conclure que l'indice est en baisse.

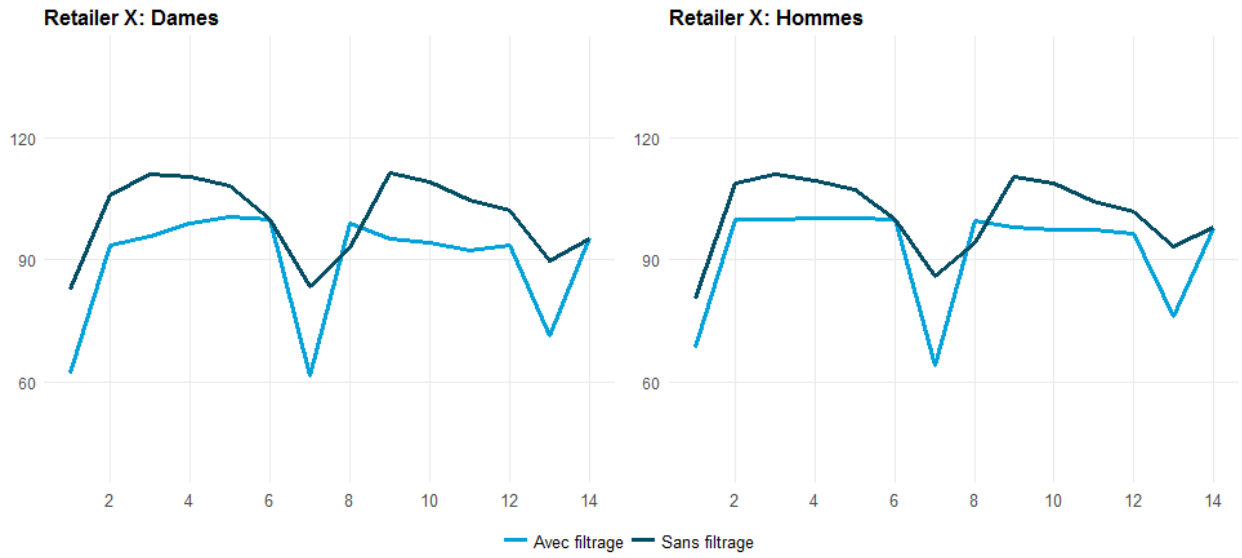


Figure 24: Calcul de l'indice avec et sans filtrage des réductions

Si maintenant on intègre cet indice aux indices des autres sites web, on obtient des indices assez similaires pour le *webscraping* et les relevés de prix classiques.

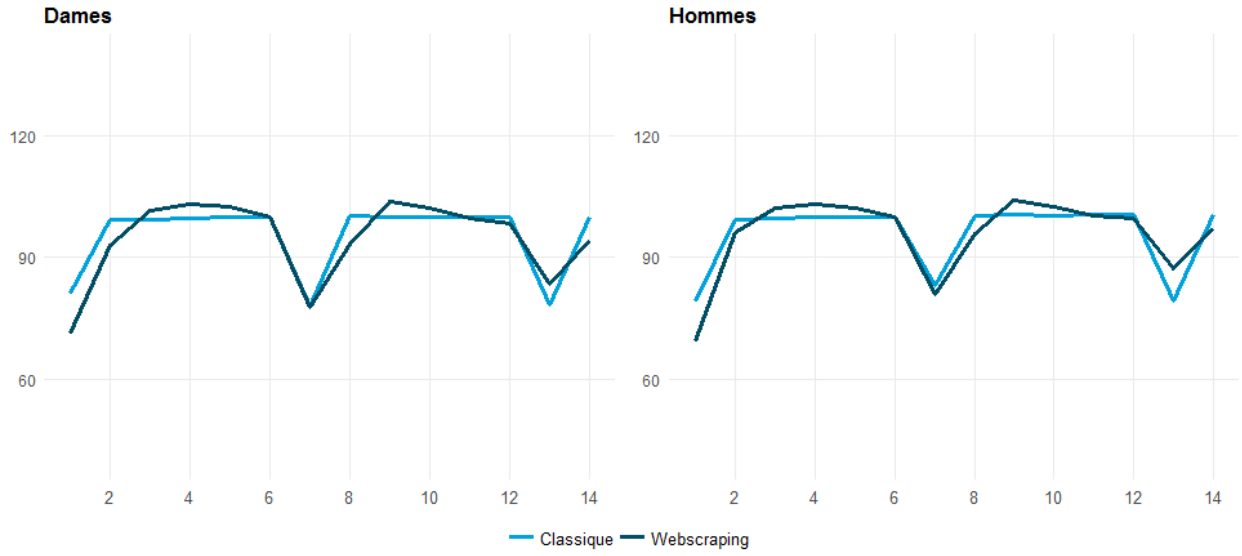


Figure 25: Comparaison des indices entre la méthode classique et le *webscraping* avec filtrage

Dans le calcul ci-dessus, seuls les "types de chaussures" ou segments de consommation disponibles pendant toute la période sont utilisés.

Il va de soi que les observations se perdent plus le nombre de périodes d'observation est élevé (parce que certains types de chaussures ne sont plus offerts pendant toute la série chronologique). De même, un nouveau type de chaussure peut arriver sur le marché et celui-ci ne serait jamais pris en compte dans l'indice. Il est donc conseillé, dans ce cas également, de travailler avec une période mobile (*rolling window*). Étant donné que les données ne sont disponibles que pour une période de 14 mois, une simulation avec une période d'estimation standard de 13 mois donnerait des résultats peu probants, parce qu'une différence ne pourrait être observée qu'au cours du 14^{e} mois. Pour ces raisons, une simulation a été réalisée avec une période plus courte, à savoir 8 mois.

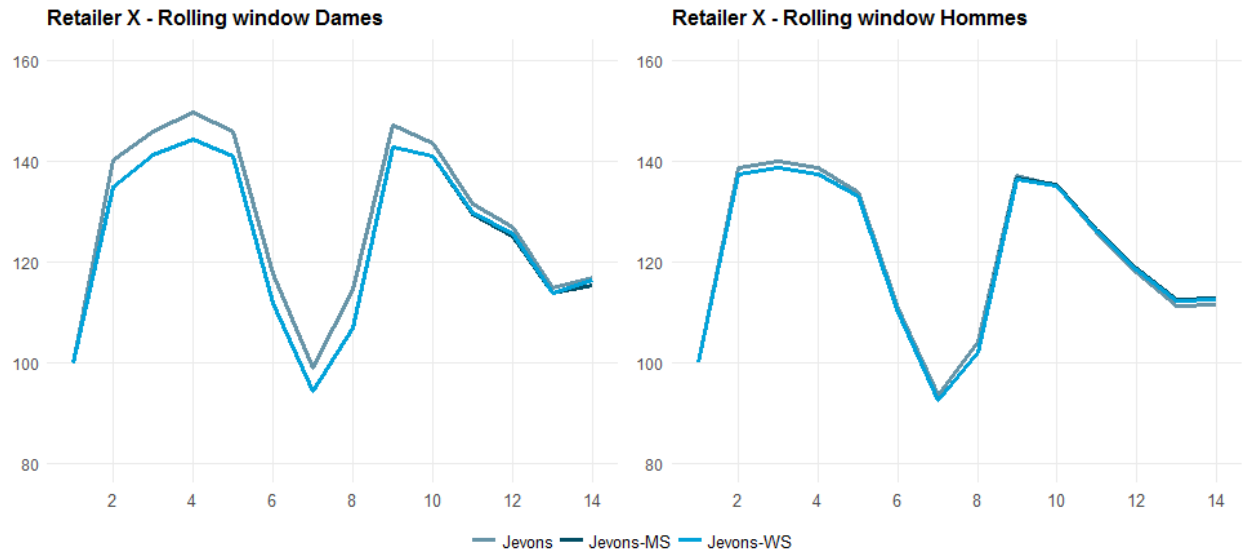


Figure 26: Comparaison entre l'indice de Jevons standard, l'indice de Jevons avec *Movement Splice* et l'indice de Jevons avec *Window Splice* [tous deux avec une période mobile (*rolling window*) de 8 mois]

Le graphique ci-dessus compare un *movement splice* (Jevons-ms) et un *window splice* (Jevons-ws), par rapport à l'indice de Jevons standard expliqué ci-dessus. La différence est minime, et l'écart entre les deux options de *splicing* est presque invisible. Dans la pratique, la période doit naturellement être plus longue que 8 mois, car une période courte risque d'inclure dans le calcul certains types de chaussures qui ne sont disponibles que pour un bref laps de temps, si bien que la différence entre le prix de lancement et le prix final serait à chaque fois prise en compte (avec pour conséquence une dérive négative compte tenu de l'absence de "*bounce-back*").

Le *window splice* est probablement préférable au *movement splice*. La méthode du *window splice* tient compte de l'évolution complète pendant cette période des "nouveaux" groupes de produits, qui sont disponibles pendant un nombre suffisant de mois et qui sont repris pour la première fois dans le calcul. En revanche, dans le cadre de la méthode du *movement splice*, seul le dernier mouvement entre périodes a une influence.

Un indice de Jevons est donc utilisé au niveau le plus bas, soit un indice non pondéré. Naturellement, aucune information relative au chiffre d'affaires ne peut être obtenue en ligne par *webscraping*. Sur la base d'un certain nombre de caractéristiques, nous pouvons identifier des articles similaires. Chessa et Griffioen (2017)[14] sont parvenus à la conclusion suivante: si l'on considère le nombre d'articles comme un proxy du nombre de pièces vendues et qu'ensuite le chiffre d'affaires est calculé en multipliant ce nombre par le prix moyen de ces articles, on obtient des résultats similaires à un indice calculé à l'aide des données de scanning du même site web. Leur exercice était toutefois basé sur un seul site web et est donc difficile à généraliser.

Nous avons tenté de reproduire cet exercice pour 1 détaillant et de calculer un indice qui tienne compte du nombre de "pièces vendues" (avec le même proxy que celui décrit précédemment), et ensuite de comparer l'indice résultant à un indice de Jevons non pondéré. La méthode utilisée pour tenir compte du nombre de "pièces vendues" est la *Quality adjusted unit value - Geary-Khamis methode* (QU-GK).

L'indice QU-GK (Chessa 2016)[15] est défini comme suit:

$$P_{QU}^{0,t} = \frac{\sum_{i \in G_t} p_i^t \, q_i^t / \sum_{i \in G_0} p_i^0 q_i^0}{\sum_{i \in G_t} v_i q_i^t / \sum_{i \in G_0} v_i q_i^0} \qquad \text{Équ. 39}$$

dans laquelle p est le prix et q le nombre de pièces vendues. Le numérateur donne donc l'évolution de la valeur totale de l'ensemble des observations pour un segment de consommation donné. Le dénominateur donne l'évolution entre la période 0 et la période t pour un nombre pondéré de transactions de chaque article dans le segment de consommation i.

[14] Chessa, T., Griffioen, R. (2017), Weights for web scraped data, Comparing scanner data and web scraped data, CBS (non publié)
[15] Chessa, T. (2016), A new methodology for processing scanner data in the Dutch CPI, Eurostat Review on National Accounts and Macroeconomic Indicators, 49-69.

L'indice est donc en fait égal à un indice de valeur divisé par un indice de quantité, ce dernier corrigé des "différences de qualité" entre produits. Le calcul de v_i n'est pas évident:

$$v_i = \sum_{z \in T} \varphi_i^z \frac{p_i^z}{P_{QU}^{0,z}} \qquad \text{Équ. 40}$$

et

$$\varphi_i^z = \frac{q_i^z}{\sum_{s \in T} q_i^s} \qquad \text{Équ. 41}$$

P_{QU} est donc l'un des facteurs qui détermine v_i, alors que v_i contribue également à déterminer P_{QU}. Les deux équations doivent donc être calculées simultanément. Une procédure itérative peut permettre la convergence. Celle-ci s'obtient en prenant comme base un indice fictif (par exemple 1).

Le graphique ci-dessous compare l'indice QU-GK résultant et l'indice de Jevons. La différence entre les deux est négligeable.

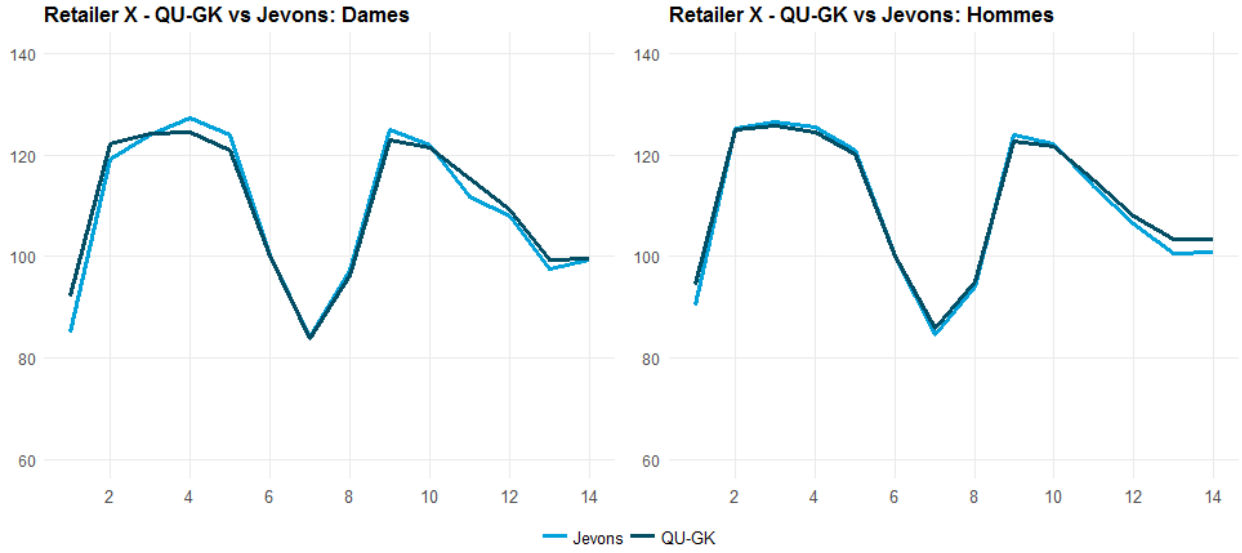


Figure 27: Comparaison de l'indice de Jevons et de l'indice QU-GK

Étant donné que la définition des articles similaires prend beaucoup de temps et que l'hypothèse selon laquelle le nombre d'articles similaires est considéré comme proxy du nombre d'articles vendus n'est pas toujours correcte, il est peut-être plus pragmatique de travailler avec un indice de Jevons non pondéré.

Entre-temps, encore plus de données ont été collectées par webscraping, et cette étude de cas sera examinée de manière plus approfondie.

10. APPRENTISSAGE AUTOMATIQUE (*MACHINE LEARNING*)

Étant donné que le *webscraping* permet de récolter de très nombreuses données, il est également possible que certaines d'entre elles soient moins intéressantes pour un traitement ou des calculs ultérieurs. Les données récoltées peuvent être filtrées et classées à l'aide de l'apprentissage automatique et d'algorithmes, de sorte qu'il ne reste que les données présentant un intérêt. Il est également nécessaire de pouvoir classer les données collectées dans des catégories générales afin de pouvoir calculer les indices par catégorie.

Le *scraping* des DVD en est un exemple simple: Un *scraping* est effectué sur certains sites web pour un certain nombre de données (titre, description, prix, etc.) relatives aux DVD disponibles. Il est toutefois possible que seuls les DVD individuels (produits homogènes) présentent un intérêt aux fins du calcul de l'indice. Les coffrets de DVD, les éditions spéciales, etc. sont donc exclues. Dans ce cas, un algorithme de classification (apprentissage automatique) peut être appliqué afin de ne garder que les données "intéressantes". On examine ensuite la description des DVD. Sur la base des mots qui apparaissent dans cette description, on peut décider de la catégorie dont le DVD fait partie (classification en fonction du texte). Bien entendu, cette classification pourrait aussi en théorie être programmée entièrement manuellement. Par exemple, grâce à la tenue à jour d'un certain nombre de "dictionnaires", lorsqu'un (ou plusieurs) mots apparaissent dans les dictionnaires préprogrammés, un article peut être assigné à une certaine catégorie sur la base de règles de décision. Cela demande naturellement de nombreux travaux de recherche manuels. De même, toute modification des dictionnaires doit à nouveau être contrôlée au niveau de la cohérence avec les classements précédents. Par ailleurs, les dictionnaires et règles de décision doivent à nouveau être adaptées à chaque nouvelle observation. Tout cela demande beaucoup de temps. Grâce à l'apprentissage automatique, tout ceci peut en grande partie être automatisé.

Afin de pouvoir appliquer l'apprentissage automatique, il convient de décider des caractéristiques que le modèle doit utiliser pour procéder à la classification. La classification des articles revient à comparer de nouveaux articles aux articles déjà classés. Les données doivent être traitées de manière à ce que les articles puissent être comparés. Une manière de procéder consiste à convertir toutes les caractéristiques en valeurs numériques. Pour la classification en fonction du texte, on utilise souvent la fréquence des mots dans les descriptions des données. La "valeur numérique" d'un mot correspond donc à la fréquence de ce mot. Il existe différents algorithmes d'apprentissage automatique qui peuvent réaliser la classification en fonction du texte, notamment la méthode des k plus proches voisins (*K-nearest neighbours*) (KNN), la classification naïve bayésienne (*naive bayes classifier*), les forêts aléatoires (*random forest*) et les machines à vecteurs de support (*support vector machine*). Chaque algorithme utilise une manière différente de comparer les articles. Ces algorithmes sont expliqués plus en détails ci-dessous.

De manière générale, on établit en premier lieu un dataset d'apprentissage, dans lequel les données sont classées manuellement. À l'aide de ce dataset d'apprentissage, chaque algorithme peut construire un modèle pour classifier de nouvelles données en fonction des similitudes entre le dataset d'apprentissage et les nouvelles données. Chaque algorithme est ensuite testé sur un dataset de test. On peut alors décider quel algorithme donne le meilleur résultat.

Avant d'appliquer l'algorithme sur de nouvelles données, ces dernières doivent encore être "nettoyées". Cela signifie que les données subissent certaines transformations afin de faciliter le traitement et l'analyse ultérieurs. Ces transformations sont aussi nécessaires pour traiter toutes les données de manière égale. Ainsi, un nouvel article peut toujours être comparé de la même manière aux articles du dataset d'apprentissage. Grâce à ces transformations, de longues descriptions sont réduites aux mots les plus importants de la description. Voici des exemples de "nettoyage":

- La conversion en caractères minuscules: pour veiller à ce que toutes les descriptions soient lues de la même manière, il est pratique de toutes les transformer en caractères minuscules.
- La suppression des mots vides: les mots vides sont des mots qui sont souvent utilisés, mais qui n'ont jamais beaucoup de sens (par exemple les articles, les conjonctions, etc.). Ces mots n'apportent aucune plus-value à la classification et sont donc souvent supprimés afin d'améliorer le fonctionnement du modèle.
- La suppression des mots communs: les mots communs ne sont pas importants lors de la classification dans les différentes catégories. Par exemple, pour le *scraping* des DVD, le mot "DVD" de la description n'est en soi pas important, étant donné que tous les éléments issus du *scraping* sont des DVD.
- La suppression des signes de ponctuation et des espaces: ceux-ci n'apportent aucune plus-value au classement.

- Réduire les mots à leur racine (*stemming*): par exemple pour "saisons", la racine est "saison".

Après application de l'algorithme, les nouvelles données classées peuvent être ajoutées aux données d'apprentissage, rendant ainsi le modèle toujours plus intelligent et plus fiable.

10.1. KNN

La méthode KMN est la méthode des k plus proches voisins (*K-nearest neighbours*). La méthode des k plus proches voisins consiste à rechercher, pour chaque nouvel article, les k plus proches voisins (d'apprentissage) dans les données d'apprentissage. Le nouvel article est alors classé dans la même catégorie que la majorité des "k plus proches voisins". Si k=1, on cherchera pour le nouvel article l'article le plus proche dans le dataset d'apprentissage, et le classement de ce "voisin" sera également attribué au nouvel article.

Le choix de k est déterminé de manière itérative et la valeur optimale est choisie et appliquée en fonction de la précision. Si seulement deux classements sont possibles, k doit logiquement être impair, afin d'exclure une "égalité". Si une égalité apparaît pour des valeurs plus élevées de k, l'algorithme choisit de manière arbitraire la catégorie attribuée.

Un "voisin le plus proche" renvoie à une notion de distance. Il convient donc de déterminer une distance entre le nouvel article et les articles des données d'apprentissage. Toutes les caractéristiques d'un article peuvent être perçues comme différentes dimensions dans l'espace. La valeur des caractéristiques indique alors l'emplacement dans l'espace où se trouve l'article (les coordonnées). En ce qui concerne la classification en fonction du texte, tous les mots (=caractéristiques) représentent ici une certaine dimension spatiale et un nouvel article correspond alors aux coordonnées de tous les mots de la description dans l'espace. Tous les articles ont donc une certaine place dans l'espace et peuvent être comparés sur la base de la distance (euclidienne) qui les sépare. La classification la plus fréquente des k plus proches voisins est donc attribuée au nouvel article.

De nouvelles données inconnues, qui n'ont aucune caractéristique correspondante au sein du dataset d'apprentissage, se voient attribuer la valeur par défaut au sein des données d'apprentissage.

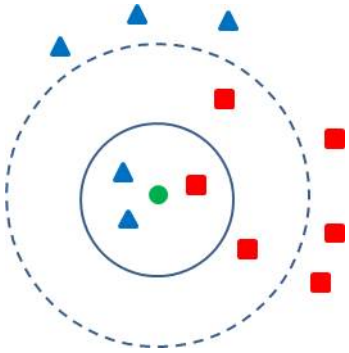


Figure 28: Visualisation de la méthode KNN

La figure ci-dessus présente un exemple de classification KNN d'articles possédant 2 caractéristiques (linéaire, coordonnées x et y). L'objectif consiste à classer le nouvel article (rond vert) soit dans le groupe des carrés rouges, soit dans le groupe des triangles bleus. Si k=3, on examine les 3 objets les plus proches (cercle intérieur). Sur la base de cet algorithme, le rond serait attribué au groupe des triangles bleus. Lorsque k=5 (cercle extérieur), la majorité des "voisins" sont classés comme "carrés rouges". Par conséquent, le rond sera classé dans cette catégorie.

10.2. Machine à vecteurs de support (*support vector machine*)

La méthode de la machine à vecteurs de support fonctionne, tout comme la méthode KNN, avec des coordonnées dans l'espace. D'un point de vue théorique, l'algorithme SVM effectue "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 le nouvel objet dans la bonne catégorie. Plus l'objet est éloigné de l'hyperplan, "plus" il appartient à cette catégorie. Il

existe différences manière de séparer les catégories. La "meilleure séparation possible" signifie que la distance entre l'hyperplan et les objets les plus proches de chaque classe (la marge) est la plus grande possible.

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

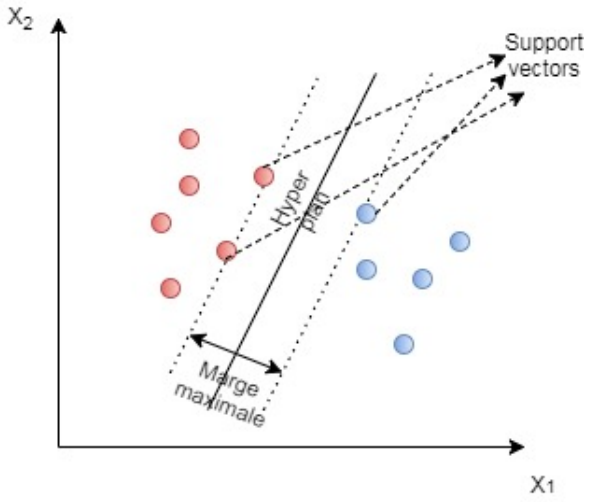


Figure 29: Présentation de la classification via l'algorithme de machine à vecteurs de support (*support vector machine*)

Les points se trouvant sur les lignes en pointillés sont appelés "vecteurs de support" parce qu'ils soutiennent l'hyperplan. La marge la plus grande possible est déterminée par les hyperplans parallèles contenant les vecteurs de support. L'algorithme de machine à vecteurs de support créera, sur la base de la description et de la classification attribuée dans les données d'apprentissage, un modèle qui attribuera de nouveaux articles à l'une des catégories.

10.3. Classification naïve bayésienne (*naive Bayes classifier*)

La classification naïve bayésienne utilise la théorie des probabilités et le théorème de Bayes pour prédire la catégorie des nouvelles données. Le théorème de Bayes dit que:

$$P(x|y) = \frac{P(y|x) * P(x)}{P(y)} \qquad \textit{Équ. 41}$$

Où : $P(x) =$ la probabilité que x se produise et $P(x|y) =$ la probabilité (conditionnelle) que x se produise étant donné que y se produit.

L'algorithme de la classification naïve bayésienne calculera pour chaque nouvel article la probabilité de chaque catégorie possible, et la catégorie avec la plus grande probabilité est attribuée au nouvel article. Le calcul de la probabilité se fonde sur la description de l'article. La formule est la suivante:

$$P(\text{catégorie 1}|\text{description A}) = \frac{P(\text{description A}|\text{catégorie 1}) * P(\text{catégorie 1})}{P(\text{description A})} \qquad \textit{Équ. 42}$$

Le terme "naïve" renvoie au fait que chaque mot de la description est pris en considération indépendamment des autres mots. Sinon, la probabilité qu'une certaine description appartienne à une catégorie serait souvent de 0, alors qu'une description exactement identique se trouve déjà dans les données d'apprentissage. En supposant l'indépendance, la probabilité d'une description est égale au produit de toutes les probabilités de tous les mots de la description. Ces mots individuels sont plus susceptibles d'apparaître dans les données d'apprentissage. Par conséquent, la probabilité sera différente de 0. Mathématiquement, l'indépendance est notée comme suit:

$$P(\text{description A}|\text{catégorie 1}) = \prod_i P(A_i|\text{catégorie 1}), \qquad \text{Équ. 43}$$

où A_i représente les différents mots de la description A.

Grâce à l'indépendance, l'ordre des mots dans une description n'a pas d'importance. La classification naïve bayésienne examine donc toujours la fréquence d'un mot dans les données d'apprentissage et la catégorie correspondante. Il suffit donc de compter le nombre de fois qu'un certain mot apparaît dans une certaine catégorie.

L'application de ces formules permet de calculer la probabilité pour les nouvelles données:

$$P(\text{catégorie 1}\,|\text{description A}) = \frac{P(\text{description A}\,|\text{catégorie 1}) * P(\text{catégorie 1})}{P(\text{description A})}$$

$$= \frac{\prod_i P(A_i|\text{catégorie 1}) * P(\text{catégorie 1})}{P(\text{description A})} \qquad \text{Équ.44}$$

Cela se produit pour chaque catégorie possible:

$$P(\text{catégorie 2}\,|\text{description A}) = \frac{\prod_i P(A_i\,|\text{catégorie 2}) * P(\text{catégorie 2})}{P(\text{description A})} \qquad \text{Équ. 45}$$

Où la description A est composée des mots A1, A2, A3, etc. La catégorie ayant la probabilité la plus élevée est finalement attribuée aux nouvelles données.

Cependant, si un mot n'apparaît pas encore dans les données d'apprentissage, sa probabilité sera de 0 (p. ex. $P(A_1|cat\ 1) = 0$), et aucun résultat ne sera obtenu. Ce problème peut éventuellement être contourné par une correction de Laplace, qui est une correction possible:

$$P(A_i \mid \text{cat } 1) = \frac{\#\ A_i + 1}{\#w_i \text{ cat } 1 + \#\ w_iT} \qquad \text{Équ. 46}$$

Où $\#\ A_i$ est la fréquence du mot A_i, $\#w_i$ cat 1 le nombre de mots dans la catégorie 1 et $\#\ w_iT$ le nombre de mots dans l'ensemble des données d'apprentissage. Il est également possible de ne pas tenir compte des mots "inconnus". Leur probabilité est alors considérée comme inexistante. Comme les nouvelles données s'ajoutent toujours au dataset complet, ce dernier s'étend constamment et il sera possible à l'avenir de tenir compte des nouveaux mots.

10.4. Forêts aléatoires (*Random Forests*)

Les forêts aléatoires (*random forests*) constituent un ensemble d'arbres de décision (*decision trees*). Dans le contexte de la classification, le résultat d'un tel arbre de décision est une certaine catégorie. Chaque arbre de décision prédit la catégorie à l'aide de certaines règles de décision. Ces règles sont établies sur la base des données d'apprentissage dont la description et la catégorie sont connues. L'enchaînement des règles et des résultats donne lieu à une structure en arbre, et une décision doit être prise à chaque nœud. En fonction de la décision, un chemin est pris en direction du résultat (catégorie). Cet ensemble de tous les arbres de décision possibles s'appelle la forêt aléatoire (*Random Forest*). La prédiction finale de la catégorie ressort de la combinaison de tous les résultats des arbres de décision. La classification comptant la majorité de voix des arbres de décision l'emporte. Normalement, quelque 500 arbres de décision sont créés. Au fur et à mesure que le nombre d'arbres augmente, l'erreur *out of bag* (pourcentage de classification erronée) diminue. Dans la pratique, après une centaine d'arbres, ce pourcentage diminue souvent déjà très fortement. La probabilité d'une classification correcte d'une valeur peut également être calculée de manière simple. Si 400 arbres classent un élément dans la catégorie A et 100 dans la catégorie B, la probabilité estimée d'une classification correcte par le modèle est alors de 80 %.

Dans une forêt aléatoire, les arbres décisionnels sont toujours créés sur la base de différents échantillons des données d'apprentissage (*bootstrapping with replacement).* Chaque échantillon, qui correspond à 63,2 % des observations dans le test

d'apprentissage complet, est utilisé comme set d'apprentissage pour la création d'un arbre décisionnel. La ramification à chaque nœud est la meilleure ramification possible dans un sous-ensemble de caractéristiques (mots). La taille de ce sous-ensemble de caractéristiques reste stable pour le même arbre décisionnel. La taille finale du sous-ensemble retenue à chaque nœud est déterminée en exécutant plusieurs fois le modèle de forêt aléatoire avec une taille différente, puis en retenant celle avec la plus faible erreur *out-of-bag*. Chaque arbre de décision prend donc des décisions de classification sur la base des différentes variables (nœuds). Les 38,8 % d'observations restantes du set d'apprentissage complet qui ne sont pas utilisées pour la création d'un arbre de décision sont utilisées pour calculer l'erreur *out-of-bag* de cet arbre. L'agrégation de tous les pourcentages de classification erronée de tous les arbres donne le pourcentage global de classification erronée. Un nœud représente une certaine caractéristique et les branches suivantes représentent les différentes valeurs de cette caractéristique. Dans la classification en fonction du texte, chaque nœud utilise des mots pour décider de la direction à prendre. Un nœud correspond donc à un mot, et deux directions sont possibles: l'observation contient le mot, ou ne le contient pas. Au début de l'arbre de décision se trouvent en principe les mots qui font la plus grande distinction. En raison du sous-ensemble arbitraire, le nombre de nœuds de chaque arbre est également différent.

La figure suivante présente un exemple simple d'arbre de décision. Le dataset sur lequel elle se base se compose de la description de plusieurs oranges (couleur et fermeté) et des catégories "mûres" et "pas mûres":

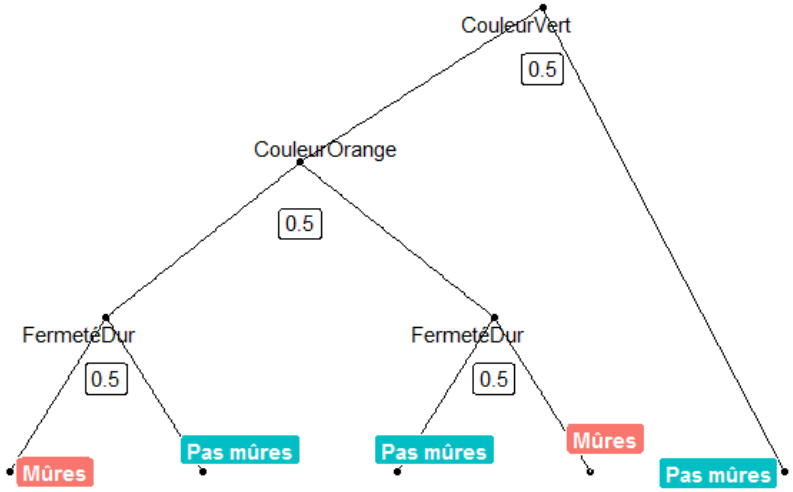


Figure 30: Exemple d'arbre de décision

10.5. Étude de cas - Les vêtements

Chaque semaine pendant 20 mois, les informations d'un site web de vêtements ont fait l'objet d'un *scraping* et ont été stockées. Ces données ont été téléchargées dans le logiciel R, où elles ont d'abord été nettoyées. Comme indiqué précédemment, cela signifie que les données sont transformées de manière à ce qu'il ne reste que les données utilisables.

Le test pratique est effectué pour deux catégories, ci-après dénommées segment 1 et segment 2. Pour ces deux catégories, on détermine les sous-catégories les plus fréquentes. Celles-ci sont alors sélectionnées à partir des données complètes. Pour les segments 1 et 2, il y avait respectivement 5 et 3 sous-catégories.

Une matrice documents-termes (*document term matrix*[16]) contenant tous les différents mots des descriptions de l'article est créée. À l'aide d'un filtre, seuls les mots pertinents (fréquence supérieure à 1) sont retenus. La matrice documents-termes est alors transformée en une trame de données (*dataframe*) sur la base de laquelle les modèles peuvent apprendre et la classification peut être testée. Les données sont divisées en un set d'apprentissage et un set de test (par exemple 85 %

[16] Une matrice documents-termes (*document term matrix*) est une matrice qui reflète la fréquence des mots qui apparaissent dans une série de descriptions. Les colonnes représentent les mots, les lignes les descriptions. Lorsqu'un mot apparaît une fois dans une description, cet élément de la matrice est égal à 1, et s'il n'apparaît pas, l'élément est égal à 0.

d'apprentissage, 15 % de test). Les deux sets de données conservent la répartition en sous-catégories telle que déterminée dans tout le dataset, sinon on pourrait obtenir des sets de données d'apprentissage et de test déséquilibrés.

Les quatre méthodes d'apprentissage automatique ont été réalisées sur ces deux sets de données. Ci-dessous figurent les résultats concernant la précision des deux segments:

Méthode	Segment 1	Segment 2
KNN	76.05%	85.28%
SVM	73.82%	83.30%
Classification naïve bayésienne	70.89%	79.53%
Forêt aléatoire	79.62%	86.13%

Il semble que la méthode de la forêt aléatoire soit la plus précise pour les deux tests. Il convient toutefois de noter que les trois premières méthodes produisent un résultat presque immédiat, contrairement à la méthode de la forêt aléatoire, qui nécessite un peu plus de temps (+/- 10 min).

10.6. Étude de cas – DVD et Blu-rays

Pour cette étude, les 100 DVD et Blu-rays les plus vendus chez un détaillant en ligne ont fait l'objet d'un *scraping* quotidien. Le dataset résultant contenait, outre les DVD et Blu-rays "individuels", également les éditions spéciales et les séries télévisées. Comme déjà indiqué au début de ce chapitre, il est important pour le calcul de l'indice d'examiner un produit plus homogène (par exemple uniquement les DVD "individuels"). Afin d'effectuer cette sélection de manière automatique, on utilise l'apprentissage automatique supervisé (*supervised machine learning*).

Environ 720 titres ont été choisis de manière arbitraire et classés dans 2 catégories: à retenir ou à exclure. Tout d'abord, les données collectées ont été nettoyées à l'aide du *text mining*. Les espaces et les signes de ponctuation sont supprimés, l'ensemble du texte est converti en caractères minuscules et certains mots sont réduits à leur élément essentiel (*stemming*). Le dataset est ensuite divisé en un set d'apprentissage et un set de test (75 % - 25 %) qui respectent la répartition entre les catégories "à retenir" ou "à exclure" dans le dataset complet.

Les 4 méthodes d'apprentissage automatique sont appliquées et les résultats sont comparés avec la classification souhaitée.

Méthode	Précision
KNN	91.95%
SVM	95.87%
Classification naïve bayésienne	92.96%
Forêt aléatoire	95.75%

Dans ce cas, c'est la méthode SVM qui obtient le meilleur résultat. La forêt aléatoire (*random forest*) offre également une précision très élevée, mais il convient à nouveau de noter qu'elle nécessite toujours un peu plus de temps que les autres méthodes.

La méthode SVM a déjà été testée sur la base d'un set d'apprentissage de 300 observations. Le modèle est alors appliqué sur 600 titres qui ne sont pas encore classés, et un indice de Jevons est calculé. Afin d'éviter une dérive négative (des DVD sortent

du marché à un prix plus bas que celui auquel ils sont entrés sur le marché), on utilise un *unmatched model*. On obtient finalement les résultats suivants:

	Blu-ray		DVD	
Mois	Avant l'apprentissage automatique	Après l'apprentissage automatique	Avant l'apprentissage automatique	Après l'apprentissage automatique
1	100	100	100	100
2	100,08	102,16	97,85	99,38
3	93,48	100,65	96,27	97,59
4	92,16	101,19	93,80	101,47

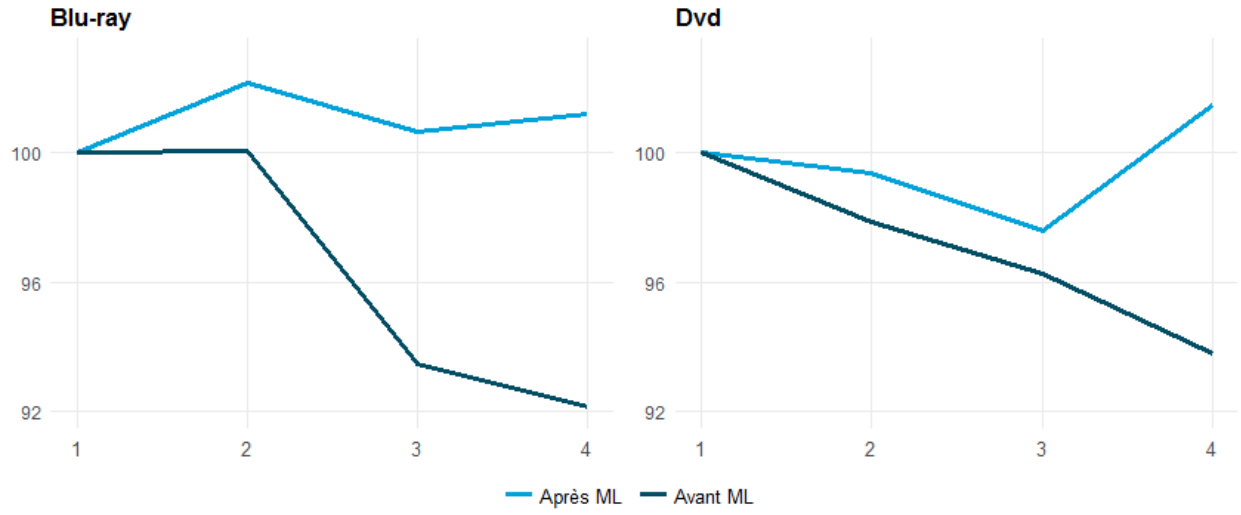


Figure 31: Indice des blu-rays et DVD avant et après l'application de l'apprentissage automatique

Avant l'application de l'apprentissage automatique, nous remarquons une diminution de l'évolution des prix. Cette baisse est due au fait qu'au début des relevés de prix, les séries et coffrets de DVD étaient également inclus dans les observations. Au fil du temps, ces observations ont disparu de la gamme, ce qui a entraîné une baisse de l'indice. L'application de l'apprentissage automatique sur le dataset complet exclut a priori les séries et les coffrets de DVD du calcul de l'indice. Par conséquent, aucune dérive négative ne se produit dans l'indice (après l'apprentissage automatique).

11. TABLEAU DE BORD

Lorsque les scripts sont enfin exécutés de manière automatique, il est bien entendu nécessaire de continuer à contrôler les résultats. Par ailleurs, les modifications des sites web doivent également être prises en compte afin de pouvoir adapter les scripts à la nouvelle structure des sites web.

Une application de tableau de bord a été conçue pour pouvoir contrôler les différents scripts de *webscraping*. De cette manière, les résultats peuvent être suivis et les scripts adaptés le cas échéant. D'une part, il convient de vérifier si le nombre d'éléments issus du *scraping* est conforme aux attentes et aux résultats précédents. D'autre part, les résultats mêmes doivent également être contrôlés. Il est donc essentiel que des prix soient collectés. Dès lors, lorsque de nombreuses valeurs sont manquantes pour cette caractéristique (prix), le site web et le script doivent être examinés.

La figure ci-dessous montre une capture d'écran du tableau de bord de *webscraping*.

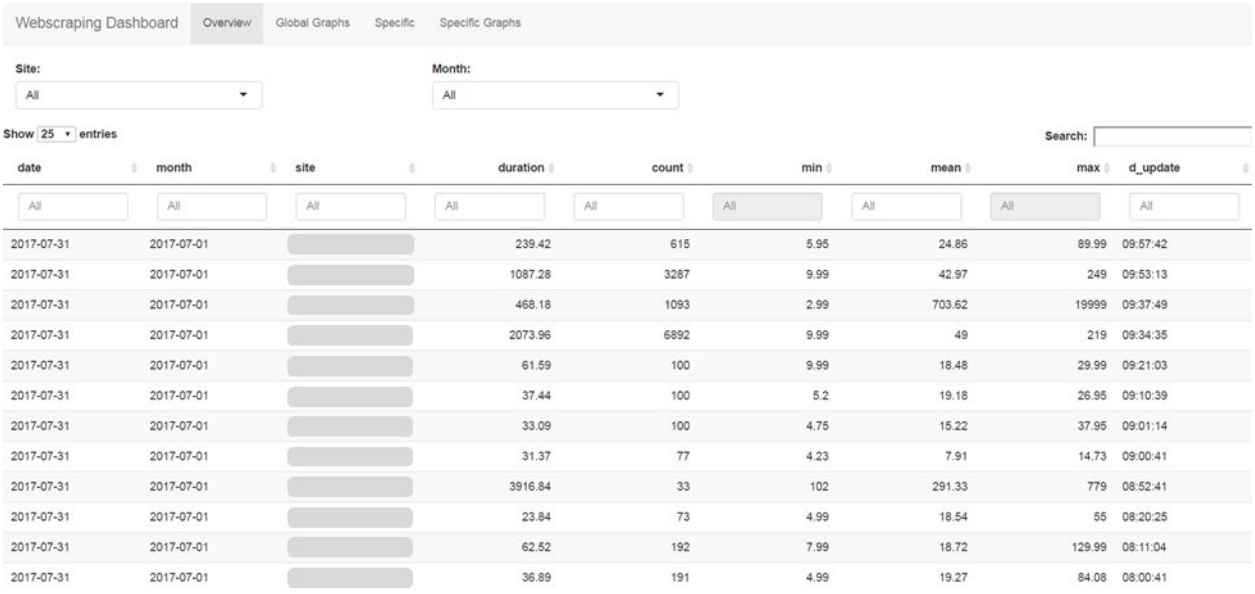


Figure 32: Tableau de bord de *webscraping* - Écran *Overview*

Le tableau de bord est composé de 4 onglets: *overview, global graphs, specific, specific graphs*. L'écran *Overview* (aperçu) affiche les scripts exécutés ainsi qu'un résumé des données obtenues par *webscraping*: date, mois, site web, durée du *scraping* (en secondes), nombre d'articles collectés, prix minimum, prix moyen, prix maximum, fin du *scraping*. Le deuxième écran *Global Graphs* (graphiques globaux) montre une représentation graphique de chaque script. La figure ci-dessous montre l'évolution de prix moyenne pour la réservation d'un séjour à l'hôtel (weekend):

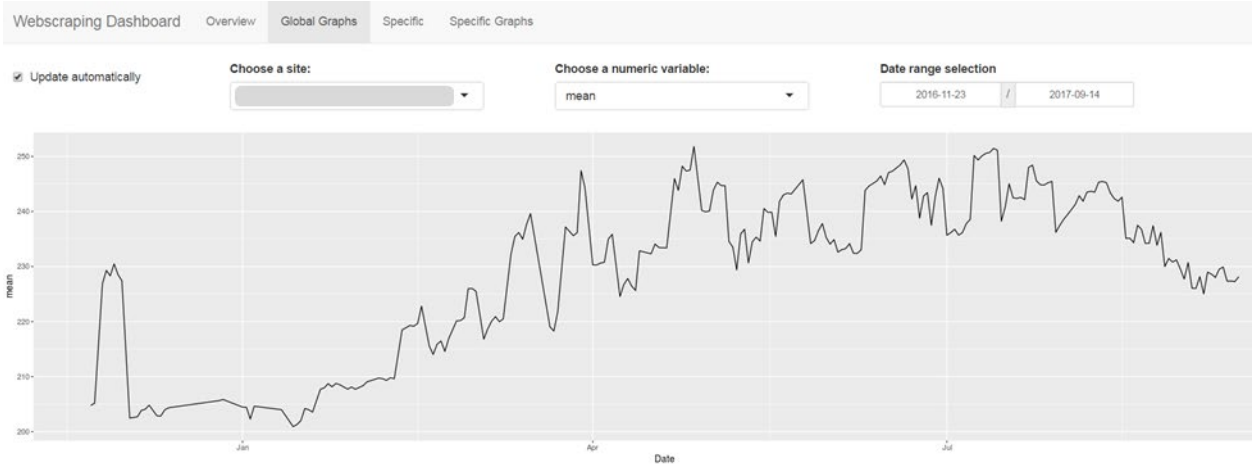
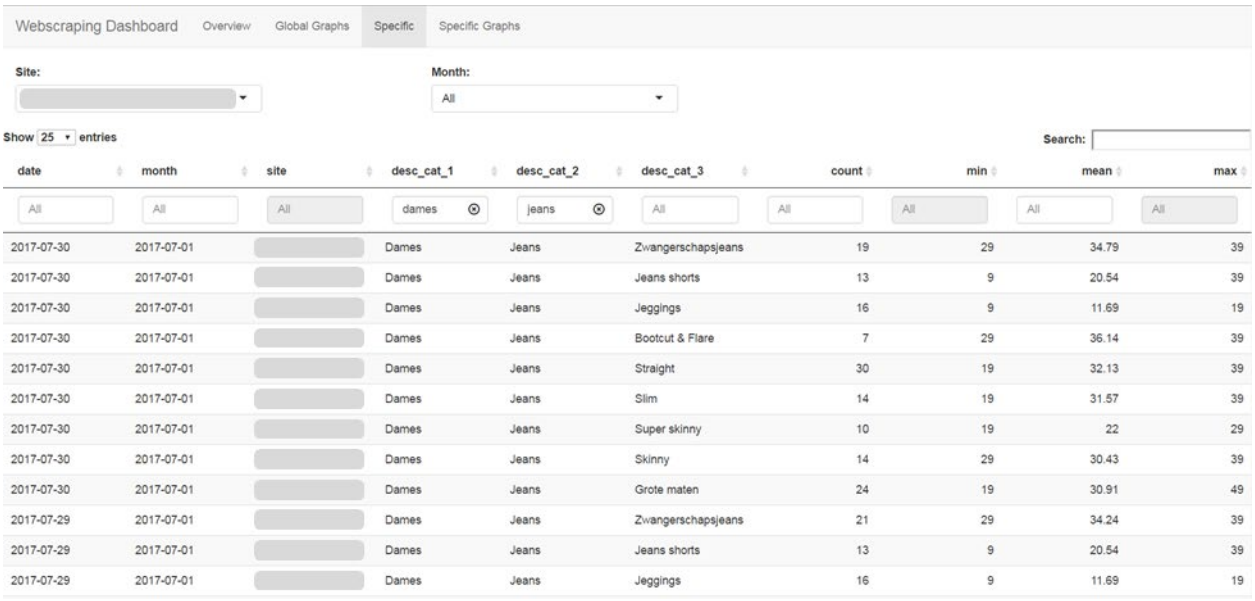


Figure 33: Tableau de bord de *webscraping* - Écran *Global Graphs*

Le troisième écran *Specific* (spécifique) présente les informations plus détaillées d'un script donné, par exemple pour les différentes sous-catégories d'un magasin de vêtements:



date	month	site	desc_cat_1	desc_cat_2	desc_cat_3	count	min	mean	max
All	All	All	dames	jeans	All	All	All	All	All
2017-07-30	2017-07-01		Dames	Jeans	Zwangerschapsjeans	19	29	34.79	39
2017-07-30	2017-07-01		Dames	Jeans	Jeans shorts	13	9	20.54	39
2017-07-30	2017-07-01		Dames	Jeans	Jeggings	16	9	11.69	19
2017-07-30	2017-07-01		Dames	Jeans	Bootcut & Flare	7	29	36.14	39
2017-07-30	2017-07-01		Dames	Jeans	Straight	30	19	32.13	39
2017-07-30	2017-07-01		Dames	Jeans	Slim	14	19	31.57	39
2017-07-30	2017-07-01		Dames	Jeans	Super skinny	10	19	22	29
2017-07-30	2017-07-01		Dames	Jeans	Skinny	14	29	30.43	39
2017-07-30	2017-07-01		Dames	Jeans	Grote maten	24	19	30.91	49
2017-07-29	2017-07-01		Dames	Jeans	Zwangerschapsjeans	21	29	34.24	39
2017-07-29	2017-07-01		Dames	Jeans	Jeans shorts	13	9	20.54	39
2017-07-29	2017-07-01		Dames	Jeans	Jeggings	16	9	11.69	19

Figure 34: Tableau de bord de webscraping - Écran *Specific*

Le quatrième écran *Specific Graphs* (graphiques spécifiques) montre une représentation graphique des résultats spécifiques:

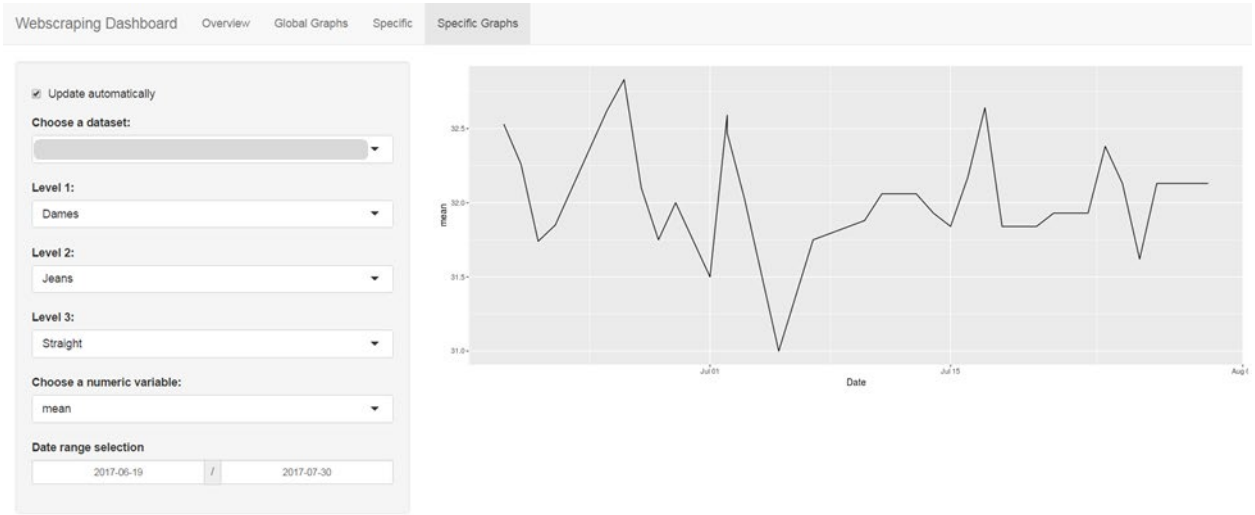


Figure 35: Tableau de bord de webscraping - Écran *Specific Graphs*

Un courrier électronique contenant les "*failed scripts*" est également envoyé tous les jours de manière automatique. À l'aide des différents logs générés lors du *scraping*, il est possible de contrôler en quoi consiste l'erreur et, le cas échéant, d'adapter le script. Les deux erreurs possibles sont:

- l'interruption de la connexion au serveur; aucune connexion ne peut donc être établie avec le site web concerné ;
- le site web est modifié (code html); le script ne fonctionne donc plus complètement.

12. ROBOTTOOL

Pour les produits qui ne changent pas beaucoup de prix ou lorsque seuls quelques prix doivent être collectés, il n'est pas conseillé d'écrire et d'exécuter des scripts qui récupèrent fréquemment les données. Non seulement cela prendrait beaucoup de temps en termes de développement du script (chaque site web est différent), mais le serveur connaîtrait également une charge supplémentaire. C'est pourquoi un outil permettant de vérifier automatiquent s'il y a eu des modifications de prix sur les sites web a été conçu. L'outil recherche un article spécifié au préalable sur le site web où le produit est présenté et est programmé pour contrôler si les caractéristiques de l'article sont modifiées par rapport au contrôle précédent. Lorsque l'outil robotisé remarque un changement, celui-ci est indiqué et l'on peut encore vérifier manuellement si la modification indiquée est correcte.

D'un point de vue technique, l'outil robotisé suit le même "*navigation path*" que celui qui aurait été utilisé en cas de recherche manuelle du produit. L'outil recherche uniquement les caractéristiques données, principalement le contexte du prix. Sinon, l'outil pourrait remarquer de nombreuses différences sur toute la page web, ce qui ne présente toutefois aucun intérêt. Mettre l'accent sur des parties spécifiques (contexte du prix) de la page web augmente l'efficacité de l'outil. Enfin, l'outil robotisé compare le contexte du prix avec le résultat de la dernière exécution. Lorsque celui-ci est différent, les différences sont montrées aux utilisateurs de l'outil (marquage). Le changement du contexte de prix n'implique toutefois pas nécessairement que le prix ait effectivement changé. L'utilisateur de l'outil robotisé peut alors contrôler ce phénomène.

La figure suivante montre le formulaire permettant d'effectuer une nouvelle vérification:

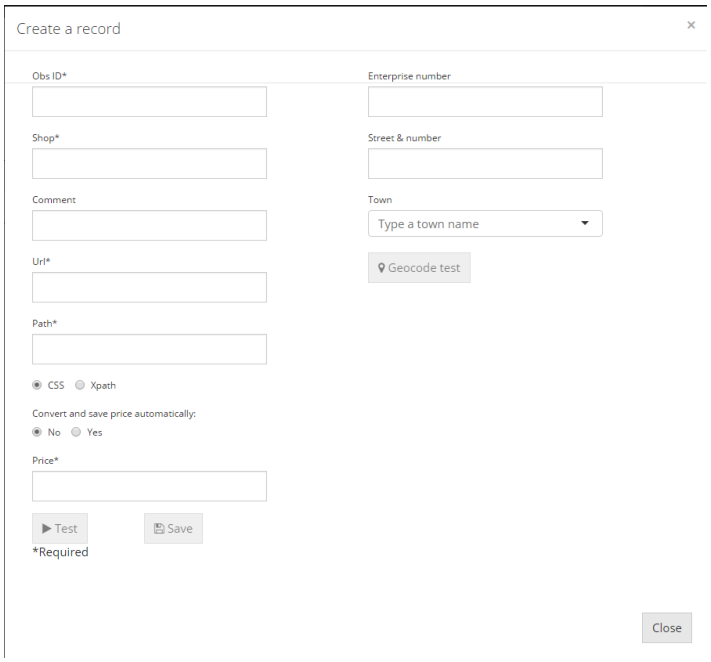


Figure 36: Formulaire permettant d'effectuer un nouveau contrôle

Lors de la création, on peut également choisir d'enregistrer automatiquement les changements de prix, le cas échéant, en tant que nouveaux prix ("*Convert and save price automatically*").

À titre d'illustration, une nouvelle vérification est créée pour un produit B. L'observation de ce produit sur le site web permet de noter un prix de 9,50 euros.

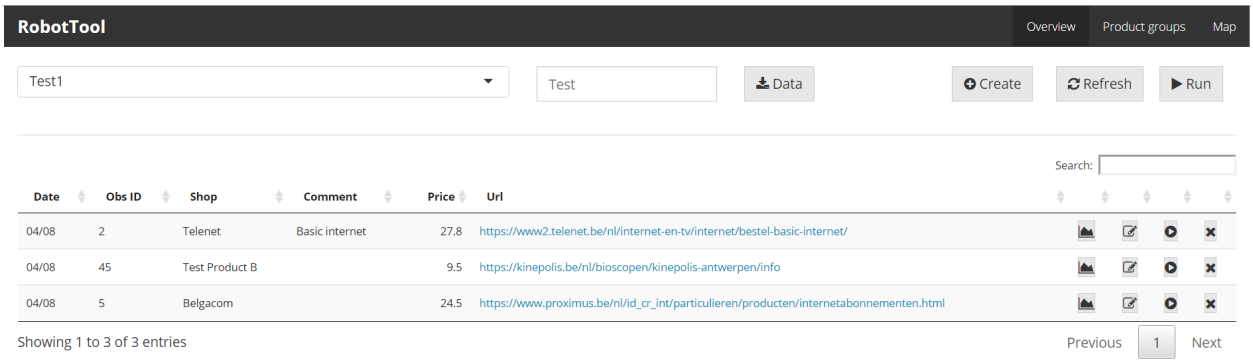


Figure 37: Exemple d'outil robotisé - Obs ID 45 Test Product B

Un peu plus tard, lors de l'exécution de l'outil, le résultat suivant est affiché:

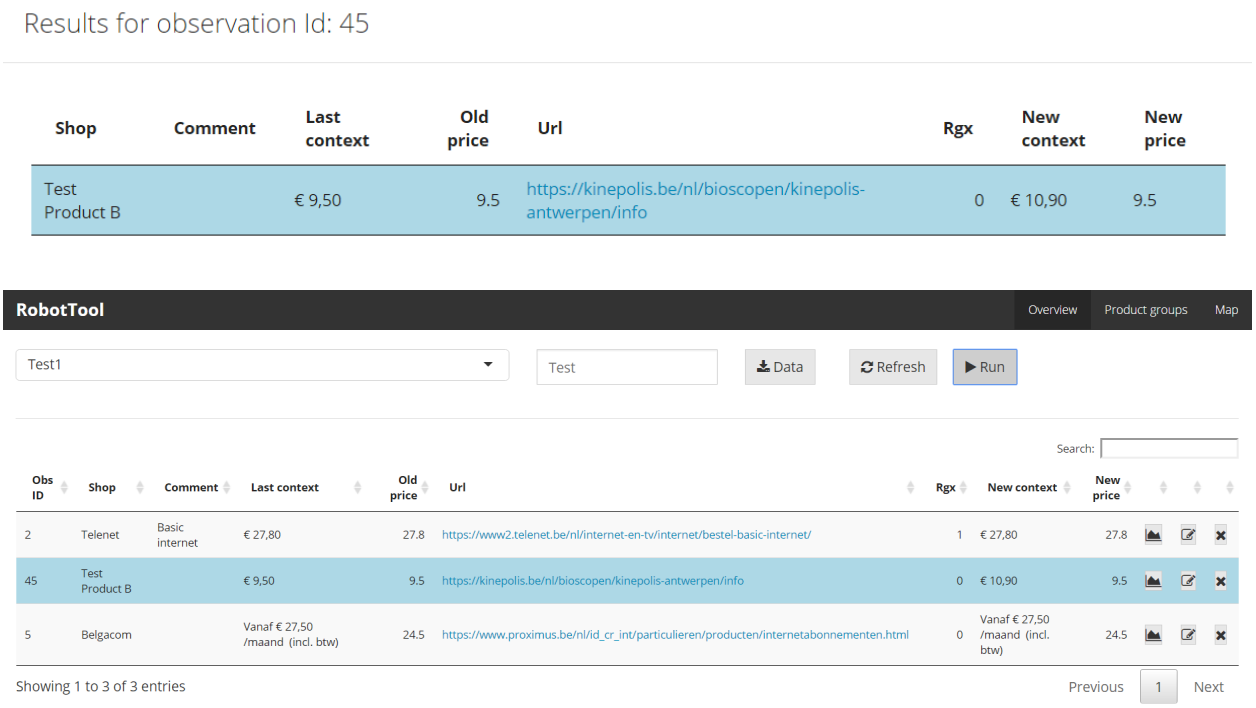


Figure 38: Marquage automatique du changement de prix

L'outil indique donc que le contexte de prix a changé. Après contrôle du produit sur le site web, on constate en effet que le prix a changé et ce nouveau prix peut être confirmé dans l'outil robotisé.

Si, lors de la création, l'option "*Convert and save price automatically*" est activée, le résultat est le suivant:

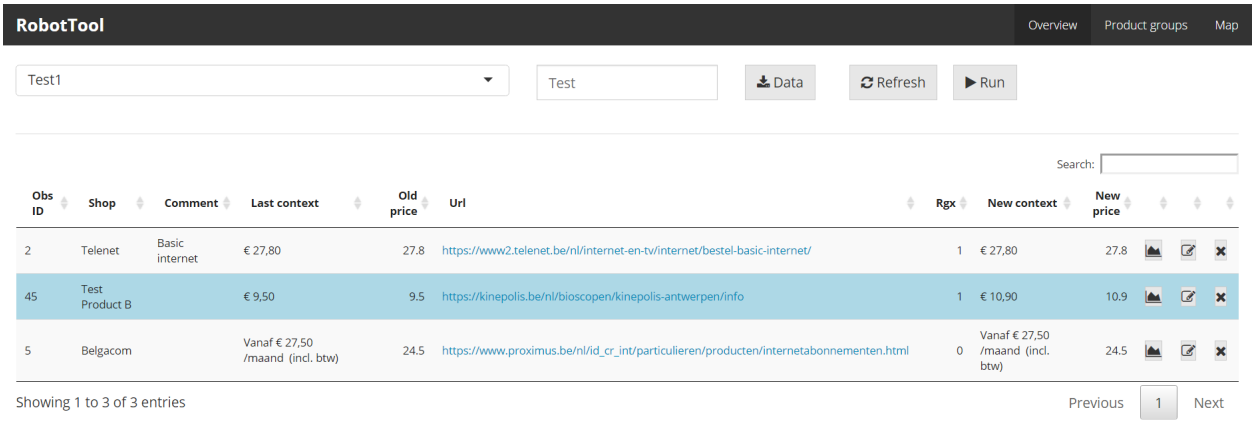


Figure 39: Marquage automatique du changement de prix avec indication du nouveau prix

Le nouveau contexte et le nouveau prix sont immédiatement affichés.

Un exemple où le contexte de prix du "*Test Product B*" a changé, mais pas le prix, est montré ci-dessous. L'ancien contexte contenait comme description "Tarif standard", alors que le nouveau contexte indique "Tarif normal". Cela est également détecté par l'outil robotisé et indiqué aux utilisateurs. Ces derniers peuvent alors éventuellement effectuer un contrôle supplémentaire.

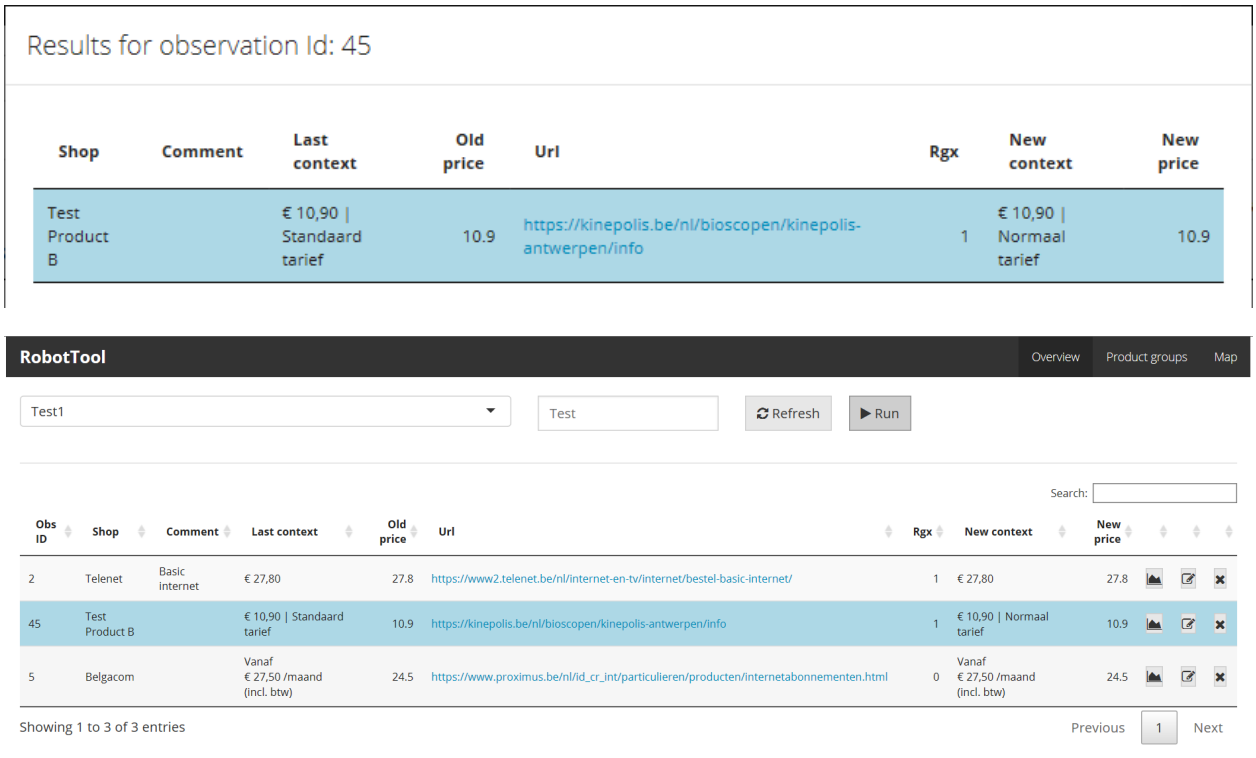


Figure 40: Marquage automatique du changement de contexte

13. CONCLUSION

Tout d'abord, on vérifie les articles pour lesquels le *webscraping* peut être testé et appliqué. Les facteurs décisifs les plus importants sont:

- les articles pour lesquels des prix sont déjà recherchés en ligne (par exemple les billets d'avion) ;
- la possibilité de remplacer la collecte de prix manuelle par le *webscraping* comme proxy des prix hors ligne (par exemple les vêtements) ;
- les articles qui sont souvent achetés en ligne (par exemple des livres, des DVD).

Actuellement, il existe des scripts pour les billets d'avion, les livres, les DVD, les jeux vidéo, un supermarché, les vêtements, les chaussures, les produits électroniques (ordinateurs portables, télévision, machines à laver, etc.), les chambres d'étudiants, les hôtels, les billets de train internationaux, les voitures d'occasion, etc.

Les résultats du supermarché, des Blu-rays, des DVD et des voyages internationaux en train sont déjà utilisés dans le calcul de l'IPC et de l'IPCH. En ce qui concerne les autres segments, des données font régulièrement l'objet d'un *scraping* et sont collectées afin d'effectuer une analyse plus poussée. Pour certains de ces segments qui sont en phase d'analyse, les résultats et les méthodes de calcul des indices sont présentés dans ce rapport à l'aide d'études de cas.

L'étude de cas relative aux chaussures a démontré que les indices des prix pouvaient être calculés au moyen du *webscraping* sur la base des données récoltées en ligne qui s'appuient fortement sur celles calculées à partir des prix relevés dans les magasins physiques. La section sur les hôtels montre qu'un échantillon peut facilement être élargi grâce au *webscraping* (plus de 1000 fois). Les études relatives aux chambres d'étudiants et aux voitures d'occasion indiquent clairement que le *webscraping* permet de capter d'éventuels nouveaux segments de consommation que l'on pouvait auparavant difficilement reprendre dans le panier de l'indice. Il ressort de la sous-partie sur les produits électroniques grand public que des méthodes hédoniques sont peut-être possibles pour un certain nombre de produits, étant donné que les caractéristiques de produit peuvent désormais également être enregistrées en combinaison avec un échantillon étendu.

Il est également apparu clairement que l'apprentissage automatique est parfois nécessaire. D'une part pour filtrer certaines observations non désirées, comme l'a démontré l'étude sur les DVD et les Blu-rays, d'autre part parce que le *webscraping* enregistre trop d'informations qu'il est impossible de classer manuellement. Dans ce cas, l'apprentissage automatique peut également être utile, comme l'a démontré l'étude sur les vêtements.

Bien entendu, l'objectif est d'élargir davantage le nombre de sites web faisant l'objet d'un *scraping*. Le suivi continu des scripts est nécessaire (par exemple lors de la modification de la structure d'un site web), c'est pourquoi le tableau de bord a été conçu. Pour les sites où seuls quelques prix doivent faire l'objet d'un *scraping*, un outil robotisé a été mis au point.

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