Index compilation with online prices for household appliances and consumer electronics

Lucien May and Botir Radjabov[1]

STATEC (Institut national de la statistique et des études économiques du Grand-Duché de Luxembourg)

Abstract:

Obtaining prices and product characteristics from the web allows statistical institutes to automatize data collection and to improve temporal coverage as well as representativeness if more products are taken into account in the Consumer Price Index (Eurostat 2020).

STATEC has been collecting prices and product characteristics from a major national e-commerce website for household appliances and consumer electronics in a bulk format for over one year using an API and expects to collect data from another similar website in the near future. STATEC is currently using only some of the collected prices by dubbing a manual price collection. The aim of this paper is to present our analysis on finding an appropriate index compilation method that exploits the full database, which includes for each month over 9000 price quotes of various household appliances and consumer electronics.

We explain how the assortment and price change dynamics let us conclude that the multilateral GEKS-Jevons method is an appropriate method for some product categories, including small household appliances, but not for all. We propose to improve representativeness of the sample by only taking into account products currently in stock and excluding those that are not in stock although customers can still order them online. We expect to introduce the multilateral GEKS-Jevons method for these products in 2024.

We also show that for some product categories, especially large consumer electronics, systematic downward trend in prices combined with high churn rates requires explicit quality adjustment methods. We test several price imputations methods, such as hedonic linear regression based and tree based price imputation methods. We conclude that random forest based price imputation method provides the most accurate results.

[1] The authors would like to thank Claude Lamboray for his useful comments and suggestions.

1. Introduction

In the last years, STATEC has been able to receive Scanner Data on a regular basis from several supermarket chains. In 2018, this new data source has been introduced in the monthly production of the National and Harmonized Consumer Price Index (CPI) for food and non-alcoholic beverages (Guerreiro et al. 2018). In 2021, the index compilation method has been changed to the multilateral GEKS method (Radjabov and Ferring 2021) and new categories, like alcoholic beverages (except wine) have since been introduced.

As part of this modernization effort, STATEC explored possible data sources for consumer electronics and household appliances. It has so far not been possible to access scanner data from the main retailers dominating the market for these products in Luxembourg. Several EU countries make use of market research data for consumer electronics (smartphones, laptops, etc.) but this type of data is currently not available in a timely manner for the Luxembourg market.

Having this in mind, STATEC has begun collecting prices and product characteristics from one of the leading national e-commerce websites (online store) for household appliances and consumer electronics in a bulk format from the beginning of 2021 on and has recently started a similar web-scraping project for the second retailer. The data covers over 9000 price quotes of various household appliances, electric appliances for personal care and consumer electronics as well as metadata on a month-by-month basis. Since online prices are known to be synchronized with the ones charged in the stores, this new data source allows to cover both online and offline purchases at once. Moreover, all products sold in store are known to also be available for online sale.

STATEC is currently using only a small subset of the collected prices by dubbing a manual price collection in a way similar to the *static* approach for processing scanner data (Guerreiro et al. 2018). That is, even though price collection is automatized, replacements of products and quality adjustments are still done manually in the office.

The aim of this paper is to present our analysis on appropriate index compilation methods that exploit the collected data in more depth and go beyond the small sample and the fixed basket approach. In particular, we will see that for some product categories it is necessary to use the collected product characteristics in order to apply explicit methods of quality adjustments.

The paper is structured as follows. Section 2 provides some general information on how the new data source has been built and on its structure. In section 3, several analyses related to the assortment are presented and in section 4, we explain that different price index methods need to be used for different product categories. Empirical results and comparisons of the different price index methods are provided in section 5, while concluding remarks are given in section 6. The annexes give supplementary information on price imputation methods and other information on selected product categories.

2. Online price collection

2.1 Introduction

In the recent years, web scraped data has become an important data source in many countries for CPI compilation, mainly due to the growing importance of online sales but also due to new methodological development for the treatment of these new data sources (Eurostat 2020). Apart from lower data collection costs, automatization of price collection offers multiple opportunities to improve the quality of the CPI such as better temporal coverage, better product coverage of the target universe (representativeness) or the collection of more metadata (product characteristics), which would otherwise not be possible.

Regarding the CPI, two types of web-scraping techniques are generally distinguished in the literature (Griffioen et al. 2016, Eurostat 2020):

- "Targeted web-scraping". The goal of this approach is to automatize price collection of specific, well-defined products by essentially mimicking a manual price collection (Blaudow, 2018). Several statistical institutes developed "robots" in order to automatize manual price collection, often for only a few prices per site, but for a large number of websites (Griffioen et al. 2016).
- "Bulk scraping" aims at collecting price and metadata for all the products from a specific website. This bulk data gives access to the total coverage of the products in a shop, but generally poses challenges in terms of validation, quality adjustment and index compilation.

Given that STATEC did not manage to get access to scanner data, the goal was to obtain "big" data of offer prices from a leading retailer by bulk scraping its website. Indeed, the scraped data includes almost the complete assortment of products on offer. However, unlike scanner data, we only have the offer prices available for online data and no information on quantity sold. This paper aims to explain how we can make the best usage of this data.

2.2 Automatic data collection

STATEC has gained first experiences in automatic online price collection by using a "point and click" scraping software, namely iMacros, to collect selected pre-defined prices on the retailer's website. This approach of "automatizing" the work of a price collector had only a limited impact in terms of improving the quality of the indices but allowed reducing data collection burden.

The efforts to gain access to scanner data (online and/or offline) of the retailer were unsuccessful, but after some discussions with the retailer it was agreed that STATEC was allowed to collect the prices of all the products on its website. Based on an informal agreement, STATEC started to scrape all price offers once per week while respecting a given crawl-delay and a given time frame (maximum one hour per night). Scraping was thus only done in the night in order to limit the impact on the servers. An internal Python script was developed by the STATEC IT department in collaboration with the price statistics unit to scrape the whole website.

The next step was to collect characteristics of several product categories in order to make it possible to apply explicit quality adjustment techniques. This task was eased by the fact that we obtained access to the API of the retailer, which facilitated the data collection process enormously. In contrast to scraping with Python by targeting the right tags in a HTML code, setting up online data collection with an API of the retailer was a matter of only a few days.

In practice, since mid-2021, bulk data collection is conducted as follows:

a) Connection to the API of the retailer via a STATEC specific User Agent using our data management and integration software.
b) Retrieval of the complete list of all product codes. A specific URL is available for this procedure.
c) Retrieval of a *.json* file (product sheet) containing all relevant information (price and metadata) for each of the products of the list (see previous step).
d) Extraction of all relevant information from the product sheets *(.json* files are subsequently deleted).

If step c) is not finished in the first cycle of data retrieval (limited to one hour), which is usually the case, it is continued 24h later until all products of the list created in the first step have been retrieved.

Having access to the API has the advantage of being less exposed to changing web site setups. Nevertheless, an internal reorganization of the retailer resulted in our access to the API to be temporarily blocked for our User Agent. Part of the data that could not be collected in the first part of the week (beginning of 2023) was exceptionally retrieved later. In order to avoid this kind of exceptional events, a formal agreement was setup between STATEC and the retailer in the beginning of 2023, which fixes the terms and conditions under which the bulk scraping is done (frequency, crawl delay, exact timing, scope, targets in the API) and also defines the range of IP addresses which STATEC uses for the data retrieval. We are confident that the bulk scraping will work in a satisfactory way for longer time periods.

In general, our experience teaches us that is important to be transparent to the retailer, which is being scraped, and to respect the netiquette (Eurostat 2020). This allows generally having a constructive relation which is essential in case of problems in the data collection process, especially because issues, if they appear, need to be resolved quickly in order to avoid loss of data. Whenever possible, access to the API of retailer simplifies the data retrieval considerably. Availability of an API should be checked upfront before starting a new scraping project.

2.3 Data structure and data preprocessing

Two separate data sets are constructed from the retrieved product sheets (*json* files). The first one concerns general information for each product including the price:

- *Reference period*: Exact date when the data has been retrieved.
- *Internal classifications:* Retailer specific hierarchical classification.

- *Unique identifiers:* Both the retailer's internal identifier and the EAN[2] code is retrieved.
- *Product name:* Unstructured label of the product.
- *Brand:* Brand of the product.
- *Prices*: Both the displayed final offer price and the strike price are retrieved (if available).
- *Discount:* Type of discount (cashback, free product, internal promotions of the retailer,…).
- *Information on the stock:* Information if the product is in stock, out of stock, available to order or available for preorder.

This database is updated every week. The retailer's internal identifier and the EAN code both give a unique product identifier. The internal classification of the retailer is very detailed[3] and a simple many-to-one mapping table allows to unambiguously classify each product into COICOP-based product classification used in the CPI. Using manual checks, we did not find evidence of misclassifications made by the retailer, which would result in false classifications in COICOP.

The second dataset is on *product characteristics* and contains a list of all product characteristics (name of the characteristic, its value, unit, creation and update date) for each product id (EAN code). We check each week if for existing products there has been a change in the product characteristics. We realized that in rare circumstances, some product characteristics can be corrected or completed one or two weeks after a product has appeared. Both datasets are created automatically at the end of the weekly process.

This raises the question of quality assurance of the bulk data sets. It is generally agreed that assuring the quality of large new (web scraped) data sources poses new challenges to official statistics (Auer, Boettcher, 2017). The price statistics unit monitors the number of products in each product category[4], the number of discount prices in each product category and checks for missing EAN codes and missing prices in the bulk data set. New internal categories appear regularly, and the mapping is thus updated every month. This being said, no major issues related to these assessments have appeared yet.

3. Analysis of the assortment

Automatization of online price collection offers in principle the possibility to monitor prices as frequently as desired and with bulk scraping the whole universe of offer prices is accessible. In comparison to traditional price collection, this allows to improve temporal coverage as well as representativeness if more products are taken into account in the CPI. Three kinds of behaviors, which are relevant in the choice of appropriate price index methods, can be analyzed:

[2] EAN - European Article Numbering are the numbers bellow the bar code of a product.
[3] Over 600 different detailed internal retailer categories are mapped to over 40 COICOP-based product classifications.
[4] In this paper, *product categories* refer to the most refined COICOP-based product classification used in the CPI.

- *Price change dynamics*: The frequency of price changes of a product in a month. E-commerce has allowed setting discounts dynamically often resulting in more volatile price evolutions (Blaudow, 2018).
- *Assortment dynamics*: The rate at which new products appear or leave the assortment over time.
- *Lifecycle pricing*: For some products, systematic price trends during the lifecycle can be observed. Thus, products can systematically enter the market with a high price and leave it with a discount. This kind of pricing strategy poses special challenges for index computation (Konny et al. 2022).

Concerning the *definition of the product*, we consider each EAN code, or equivalently internal code as a distinct "product".

The *price of a product* in a given week is the observed offer price taking into account cashbacks[5] and internal promotions of the retailer. Other promotions, like "free products", are not taken into account.

3.1 Price collection frequency

The market for consumer electronics is very competitive and it is important to know that the website we are retrieving data from is not conducting any individualized pricing. This is a pricing strategy used by online retailers by which different persons can see different prices on a website visited at the same time, depending on their location, IP address, etc.

However, we don't have information from the retailer about their price setting behavior. Therefore, an important question is to know if the frequency of data retrieval (weekly) is sufficient to give representative monthly price (changes).

For the sake of the computation of a representative monthly price, it is important to examine the share of products whose prices change twice or more per month, because a high share indicates that weekly data retrieval might not be sufficient.

Table 1: Frequency of price changes within a month (average over a period of 23 months)

Product category	Average number of products per month	Average frequency of price changes within a month (average over 23 months) (%)*		
		0	1	2 or more
Televisions	150	61.6	29.8	8.7
Dryers	32	73.9	22.2	4.0
Memory cards	93	76.0	20.8	3.3
Laptops	174	73.1	23.6	3.2
Washing machine	76	76.1	21.1	2.8
Freezer	32	85.2	13.0	1.9

[5] Cashbacks are only observed for some product categories.

Electric razor	102	88.9	9.4	1.7
Dishwasher	24	91.2	10.9	1.5
Other product categories	-	78-92	2-19	<1.5

*Based on products in stock and available at least four times in a month.

For most product categories, including small household appliances (toasters, microwaves, coffee machines, etc.), prices are changing either not within a month or at most once per month. We observe only a relatively dynamic price setting behavior for televisions, where the average share of products whose prices change two times or more is almost 10%. This high share is due to multiple retailer-specific discounts for televisions.

We define the *monthly aggregated price* p_i^t of a product i by

$$p_i^t = \frac{1}{L_{it}} \sum_{k=1}^{L_{it}} p_i^{t,k}, \qquad (1)$$

where $p_i^{t,k}$ is the offer price of product i in month t in kth-week (if available). L_{it} corresponds to the number of weeks the product i is available in month t. Given the price change dynamics described before, this price can be considered to be representative for a given product and a given month. The weekly data retrieval is clearly sufficient.

In practice, only the first three weeks can be taken into account as the CPI is computed in the last week of a given month. In section 4, we will see that indices computed by taking into account only the first three weeks of data do not differ substantially from those taking into the data for the full month.

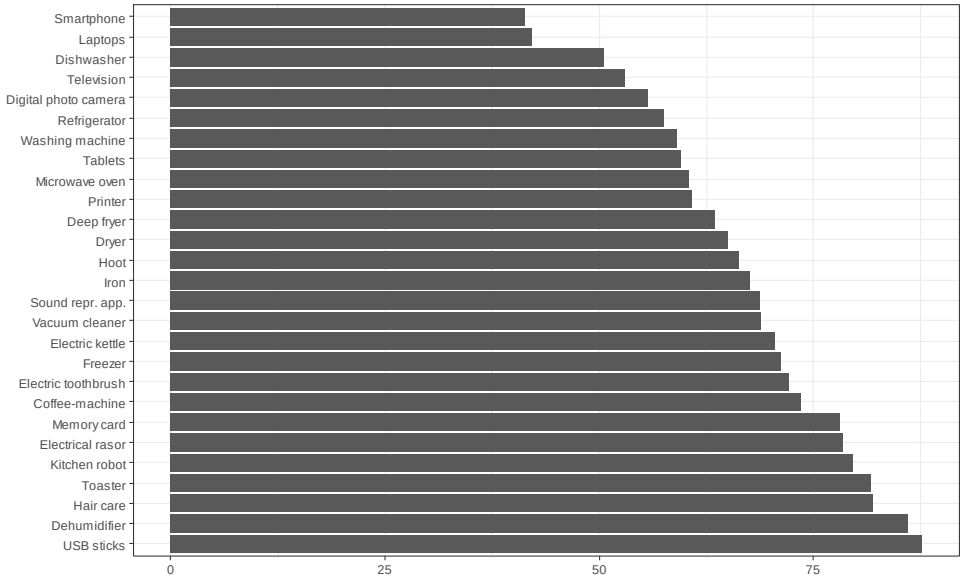
3.2 Assortment dynamics and lifecycle pricing

Assortment dynamics

The total number of products observed in each month as well as the number of products in each category has been relatively stable in the last two years. However, the assortment itself can be very dynamic: products become temporarily or permanently unavailable and new products appear. It is important to have this behavior in mind as the choice of an index method can crucially depend on this dynamics (Eurostat, 2020).

In order to analyze these dynamics, the average shares of products in terms of the number of products in a product category, which can still be matched 6 months later, have been computed. Figure 1 shows that these matching rates vary considerably between different product categories. A matching rate is defined as the number of matched products in the two comparison periods compared to the number of products available in the base period.

Figure 1: Matching rates over 6 months for selected product categories (in %, average over 23 months)



Consumer electronics (laptops, smartphones, televisions, etc.) have very high churn rates. For smartphones and laptops, on average, almost 60% of products are not available anymore after six months. Some large household appliances (washing machines, dryers) also have relatively high churn rates. Generally, we observe that small household appliances and electrical appliances for personal care have much more stable assortments. This has important consequences for the choice of an appropriate index method (see section 4).

Figure 2: Matching rates over time for selected categories (in percentage, average over 23 months)

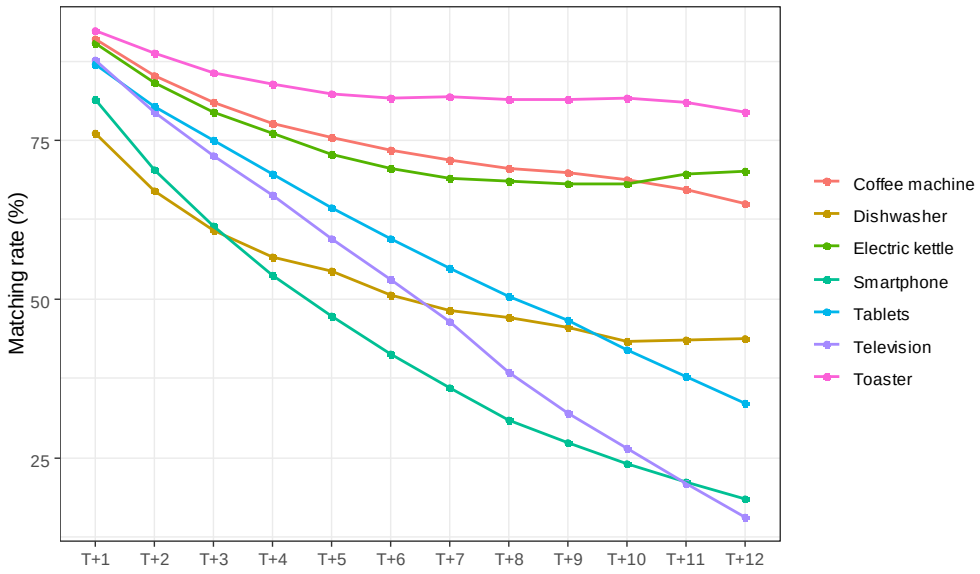
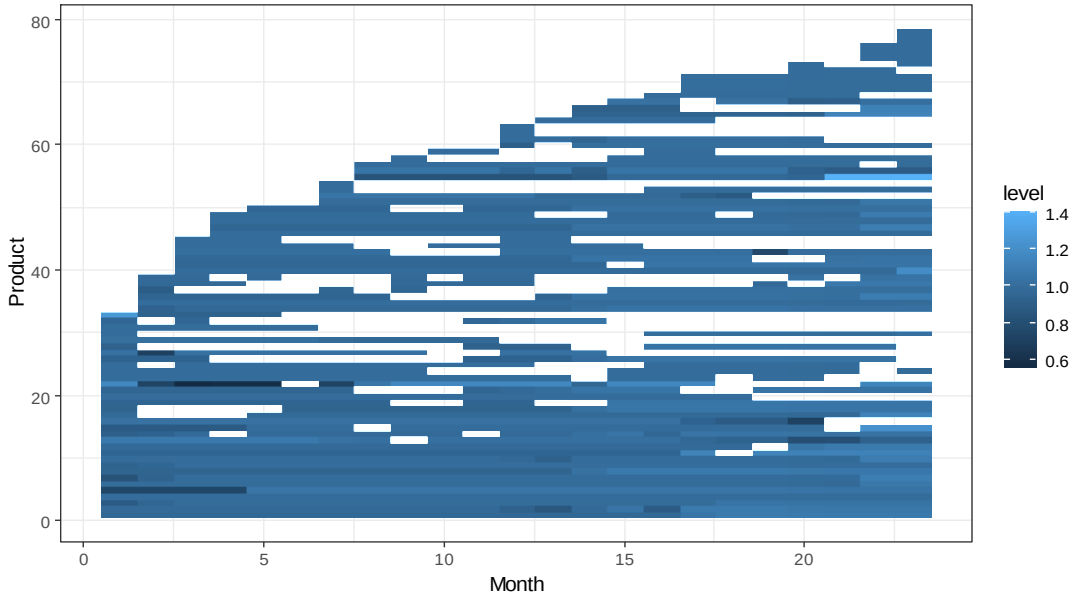


Figure 2 shows in more detail, for selected product categories, that the percentage of matched products is diminishing the higher the time interval between a reference month

and comparison month is. However, for small household appliances and electrical appliances for personal care, the matching rate after 6 month or even later remains high, making long range price comparisons based on matching like with like products possible. For consumer electronics, products are replaced at a significantly high rate resulting in very low matchings rate after 6 months or later. Note that disappearing products are largely replaced by new products, which might have different or new features as the total number of products remains relatively stable. High churn can be due to the fact that new models appear frequently on a market but it can also result from the way the retailer manages its inventories.

Even for product categories with relatively low churn, like small household appliances, we observe that many products are temporarily unavailable. Figure 3 gives an example for toasters, where each line corresponds to a product. A blank space indicates that the product is temporarily out of stock, but can reappear for sale a month later. Several products are observed during the whole period of time. It is important to choose an index method, which can deal with these temporarily missing products in an appropriate way. As we will see, this is the case with multilateral methods.

Figure 3: Dynamics of the assortment of toasters (in stock)



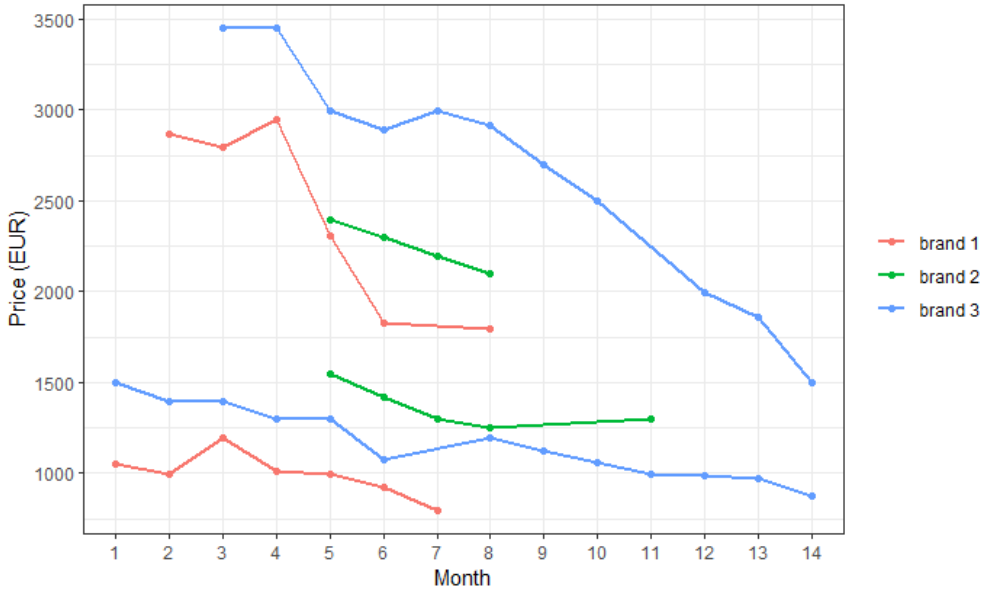
Each line corresponds to a product. The "level" is the relative price compared to the average of the whole window.

Lifecycle pricing

Apart from high churn in the assortment, another challenge for some product categories is to deal with product lifecycle effects, that is, when products exhibit systematic, usually downward price trends in their lifecycle. For certain product categories, products typically enter the market at a high price and gradually discount over time. The last observed price is then substantially discounted and may be on clearance. This kind of behavior is observed in our dataset for consumer electronics (laptops, televisions, smartphones) and poses significant challenges for index computation.

Figure 4 show typical price evolutions for some televisions in our dataset, which illustrates lifecycle pricing. Similar trends can be observed for smartphones and laptops. For these product categories, more than 90% of the products leave the sample with a discount.

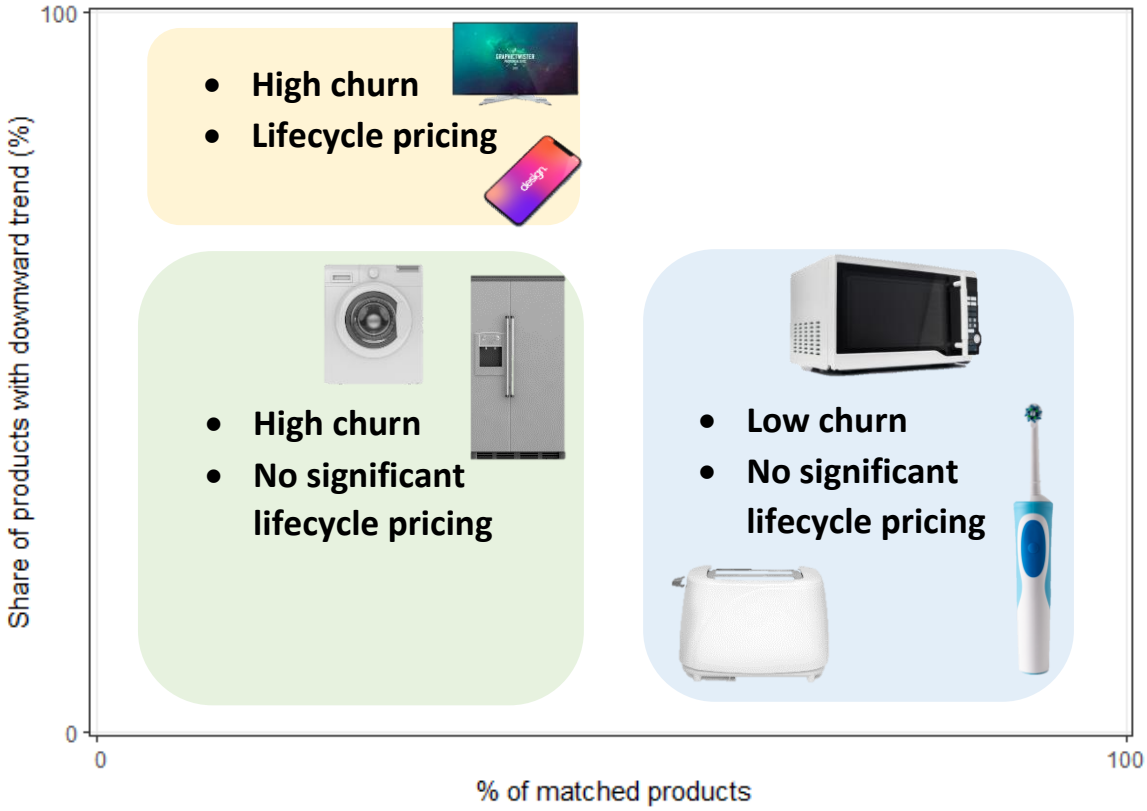
Figure 4: Typical price evolutions over time for televisions (monthly prices)



From a conceptual point of view, product categories can thus be classified according to the assortment dynamics and lifecycle behavior.

- *Stable assortments (low churn) without lifecycle pricing.*
  This is the case for the majority of the product categories, namely small household appliances, electrical appliances for personal care and some other products (scanners and printers for example).
- *Dynamic assortments (high churn) with lifecycle pricing.*
  This is the case for smartphones, televisions, laptops and tablets.
- *Dynamic assortments (high churn) without lifecycle pricing.*
  For some product categories, high churn is observed but products do not have systematic price trends. This is mainly the case for big household appliances (dishwashers, refrigerators, etc.)

Figure 5: Conceptual classification of the product categories according to assortment and lifecycle behavior



4. Price index methods

The choice of an appropriate price index method for the dataset depends on many factors.

- The rate with which products leave the assortment (and the rate with which new products appear) and systematic trends in the lifecycle of products (see section 3.2, figure 5).
- Technological developments. New products can enter the market with new or improved features. This requires quality adjustment methods to capture pure price changes.
- Weight of the considered product category.

Stable assortments (low churn) without lifecycle pricing.

In the case of stable assortments as is the case for small household appliances, electrical appliances for personal care (see section 3.2, figures 1 and 5), an unweighted matched model approach can be appropriate (Eurostat 2020). Indeed, in this case long-range comparisons of matched products take into account high shares of products. It may not be efficient to apply explicit quality adjustment methods here. All price comparisons are based on offer prices and no quantities sold are available. These methods are presented in section 4.1 and selected results are discussed in section 5.1.

Dynamic assortments (high churn) with lifecycle pricing.

In the case of assortments with high churn, simple matched models can be problematic as price indices would rely on low shares of matched products. Moreover, a systematic price downward trend during the lifecycle makes a simple unweighted matched model inappropriate as the resulting indices would be downward biased (Van Loon et al. 2018). Finally, innovative products often appear with new or improved features and quality adjustments need to be made to capture pure, constant quality price changes. Thus, for consumer electronics (laptops, tablets, televisions, smartphones), methods that use explicit quality adjustments should be used. An unweighted multilateral approach, which takes into account imputed prices of unmatched products, is presented in section 4.2 and empirical results are discussed in section 5.2.

Dynamic assortments (high churn) without lifecycle pricing.

For product categories in this case, simple matched models can be problematic as price indices would rely on low shares of matched products. One also needs to check to what degree explicit quality adjustments can influence the result and if it is worth using a more complex method (see previous case). Big household appliances (refrigerator, dishwasher, etc.) are in this category as churn rates are relatively high, but no systematic (downward) price trends can be observed. No empirical results for these product categories will be presented in this paper.

4.1 Unweighted methods, which do not use product characteristics explicitly

One of the main disadvantages of online data as compared to scanner data is the absence of individual product weights. However, several unweighted price index formulas are available. These methods will be discussed below. In this section p_t^i will denote the average price in month t of product i (see Equation (1)).

Fixed base Jevons index

Prices of individual products of the current period are compared to a fixed base month (December of previous year). The *fixed base Jevons index* corresponds to the Jevons price index in period t with 0 as base period, defined as follows:

$$I_J^{0,t} = \prod_{i \epsilon U_M^{0,t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{N_{0,t}}} \quad (2)$$

Here $U_M^{0,t}$ is the set of matching products between the base period and the reference month t. $N_{0,t}$ is the number of matched products between periods 0 and t. As we move away from the base month, the number of matched products generally decreases due to product churn. Moreover, with this method, too much emphasis is put on products in the particular base month. This method will thus not be considered.

Chained Jevons index

The *chained Jevons index* between two months is obtained by chaining monthly matched Jevons indices. It is computed as follows:

$$I^{0,t} = \prod_{j=1}^{t} I_J^{j-1,j} \tag{3}$$

where $I_J^{j-1,j}$ is the Jevons price index in period j with $j-1$ as base period as defined in equation (2) above. As a result of this monthly chaining, long-term price comparisons are not included in this index formula, which might be problematic for temporarily missing products. The next two methods provide a solution to this drawback.

Multilateral GEKS[6]-Jevons index

A *GEKS-Jevons index* between the months 0 and t, based on data between months 0 and T ($T \geq t$), is defined as follows:

$$I_{GEKS-J}^{0.t} = \prod_{j=0}^{T} \left(I_J^{0,k} I_J^{k,t}\right)^{\frac{1}{T+1}} \tag{4}$$

where $I_{s,k}^{J}$ is the Jevons price index in period s with k as base period as defined in equation (2) above. Note that the multilateral GEKS-Jevons price index includes essentially all possible price comparisons between matched products of all the periods taken into account (that is months 0 to T).

Time Product Dummy index

The *Time Product Dummy index (TPD)* uses an ordinary least squares (OLS) regression on the dataset of all products available between months 0 and T to compute a price index. The regression equation reads as follows:

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t \tag{5}$$

where δ^t are temporal parameters, γ_i product specific constants. The variable D_i^t takes the value of one, if product i is available at time t and zero otherwise, D_i is a dummy variable that has the value of one if the observation relates to product i and zero otherwise. The multilateral index is obtained directly from the model parameters:

$$I_{TPD}^{0,t} = e^{\widehat{\delta^t}}. \tag{6}$$

In section 5.1 several results will be presented, and we will explain that multilateral index methods are preferred to chained Jevons indices mainly because they seem to be less downward biased.

4.2 Explicit quality adjustment method

For product categories, which experience high churn and systematic downward price trends, simple matched model index methods (see section 4.1) are inappropriate. In order to capture pure price changes, one needs to make use of explicit quality adjustment methods, which can be done if product characteristics are available.

The Imputation Jevons GEKS (IJGEKS) method can be considered (de Haan and Diewert 2017, pp. 7-8) as a possible multilateral method, which incorporates explicit quality adjustments in form of price imputations for missing products and which is

[6] The GEKS index is named after works of Gini (1931), Eltetö and Köves (1964) as well as Szulc (1964).

directly related to the multilateral GEKS-Jevons method (see equation (4)) Therefore, the Imputation Jevons and the IJGEKS methods can be defined as follows:

$$I_{IJ}^{0,t} = \prod_{i \epsilon U_M^{0,t}} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{2N_0}+\frac{1}{2N_t}} \prod_{i \epsilon U_D^{0,t}} \left(\frac{\hat{p}_i^t}{p_i^0}\right)^{\frac{1}{2N_0}} \prod_{i \epsilon U_N^{0,t}} \left(\frac{p_i^t}{\hat{p}_i^0}\right)^{\frac{1}{2N_t}} \quad (7)$$

$$I_{IJGEKS}^{0,t} = \prod_{j=0}^{T} \left(I_{IJ}^{0,k} \Big/ I_{IJ}^{t,k}\right)^{\frac{1}{T+1}} = \prod_{j=0}^{T} \left(I_{IJ}^{0,k} I_{IJ}^{k,t}\right)^{\frac{1}{T+1}} \quad (8)$$

In Equation 7, p_i^0 denotes a price of a product i in a base month 0 and p_i^t denotes a price of a product i in a comparison month t. Moreover, $U_M^{0,t}$ denotes a subset of matched products, which are available both during a base month 0 and a comparison month t, $U_D^{0,t}$ denotes a subset of disappearing products, which are available in a base month 0 but not in a comparison month t, and $U_N^{0,t}$ denotes a subset of new products, which are available in a comparison month t but not in a base month 0. Altogether, a condition of $U^0 \cup U^t = U_M^{0,t} \cup U_D^{0,t} \cup U_N^{0,t}$ holds. Additionally, $\hat{p}_i^0$ and $\hat{p}_i^t$ denote imputed prices and N_0 denotes the number of products in the base month 0, N_t the number of products in the comparison month t.

There exist several price imputation methods for calculating imputed prices $\hat{p}_i^0$ and $\hat{p}_i^t$. However, it is not entirely clear which price imputation method should be preferred in practice. Therefore, we test several price imputation methods to find out which method provides the most accurate results. All the price imputation methods, which are tested, are provided in Table 2 below and are explained in more detail in Annex B.

Table 2: The chosen Price Imputation Methods

Hedonic Linear Regression Based Price Imputation Methods	**Tree Based Price Imputation Methods**
Time Dummy Hedonic Regression	Regression Tree
Time Dummy Hedonic Polynomial Regression	Bagging Trees
Time Dummy Hedonic Regression with Interactions	Random Forest
Time Dummy Hedonic Polynomial Regression with Interactions	

5. Empirical results

5.1 Unweighted methods, which do not use product characteristics explicitly

In the case of product categories with relatively stable assortments and in the case of the absence of individual product weights, a chained Jevons index or an unweighted multilateral index can be used to compile individual price indices. This is the case for small household appliances, electronic appliances for personal care and selected other product categories (printers, etc.), all-in-all 20 product categories (see Annex A), making up approximatively 0.7 per cent of the CPI basket. We compiled GEKS-Jevons, TPD and chained Jevons indices using all the weeks of a month in one case and only the first three weeks of every month in the other case. The window length for the GEKS-Jevons and TPD indices is 23 months[7] as this is all the data currently available. A variable is available that indicates if a product on offer is "in stock", or not. In order to increase the representativeness of the data only products "in stock" were considered. By restricting the scope to "in stock" products, we cover the most important products that consumers are more likely to purchase. For some categories up to 20% of the products are only available to order and are thus not immediately available for pickup or home delivery.

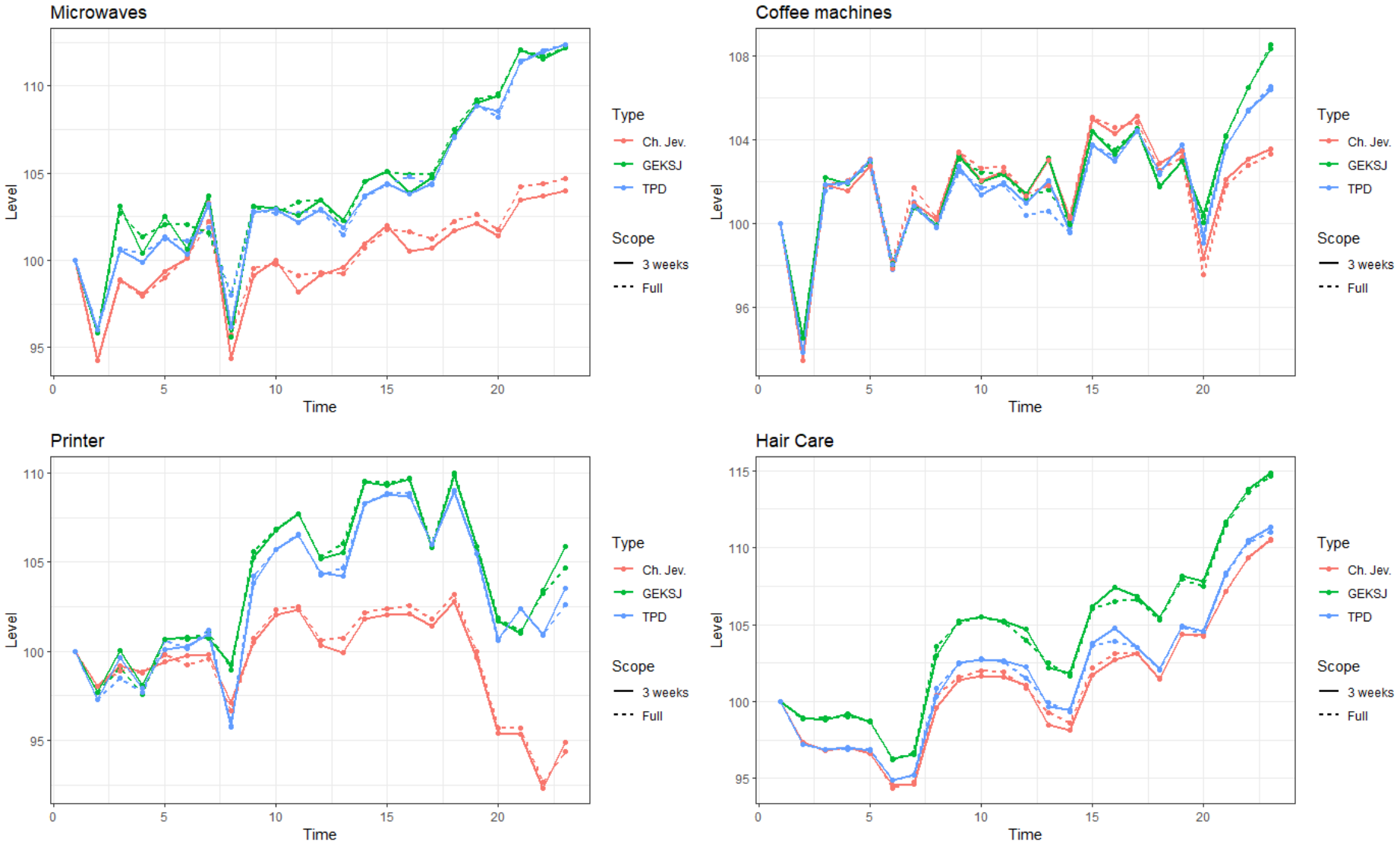
A few examples are shown in Figure 6. In all cases, the GEKS-Jevons and TPD indices are either, on average, higher than the chained Jevons indices or the difference is close to 0. This is due to the fact that the chained Jevons only compares prices of consecutive months and in case if a product is temporarily missing and reappears with a higher price, the resulting index will be downward biased. In contrast, multilateral indices integrate longer-range price comparisons and solve this problem. They are also known to be "chain-drift free".

GEKS-Jevons and TPD indices are generally similar, as it was expected (de Haan and Hendriks 2013). The TPD indices are however more sensitive to sales periods compared to the GEKS-Jevons indices (see month 8 for microwaves or printers for example in figure 6) and in some cases the TPD index stays on a slightly lower level than the GEKS indices (see for example printer and hair dryers in Figure 6), even after a sales period. These cases need to be analyzed in more detail by taking a closer look at the microdata. In general, the difference between the "three weeks" indices and the "full month" indices are very small, whatever the index type is. This indicates that producing indices in production with only the first three weeks of data is acceptable.

STATEC currently uses the "GEKS HASP" on 25 month's setup for Scanner Data. In order to be consistent with this method, a preference is given to the GEKS-Jevons method compared to the TPD approach. We expect to introduce the GEKS-Jevons index in 2024 for these low-churn product categories. This approach is compatible with the general guidelines of choosing an index method for web scraped data of Eurostat (Eurostat 2020, p. 20). Note that when a new month of data becomes available and a new multilateral window is used, multilateral indices are revised. In order to tackle this revision problem, an appropriate splicing method needs to be used. The choice of a splicing method and an optimal window length are not discussed in this paper.

[7] Data from June 2021 to April 2023.

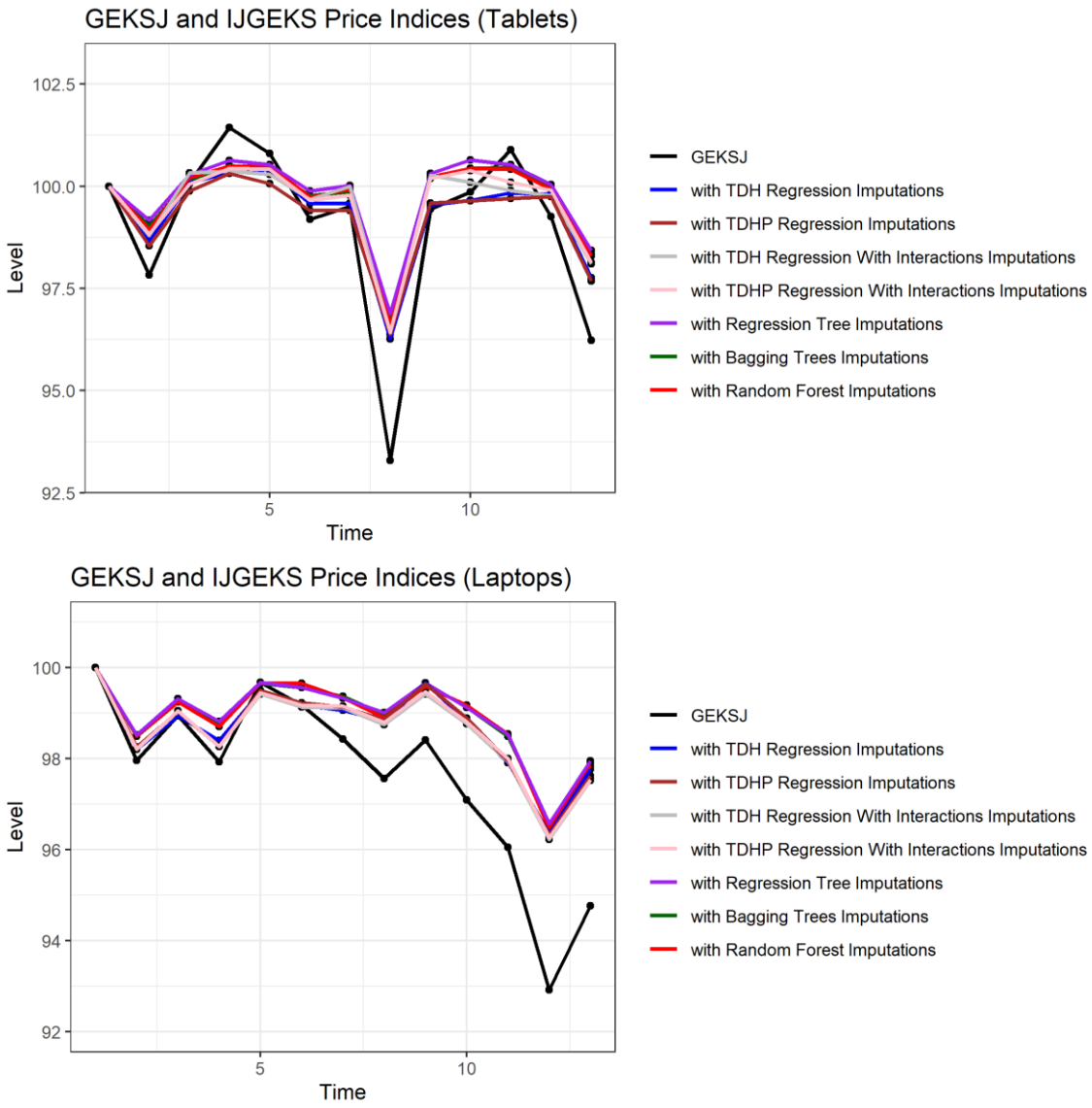
Figure 6: Comparison of chained Jevons, GEKS-Jevons and TPD indices for selected examples (full window of 23 months).

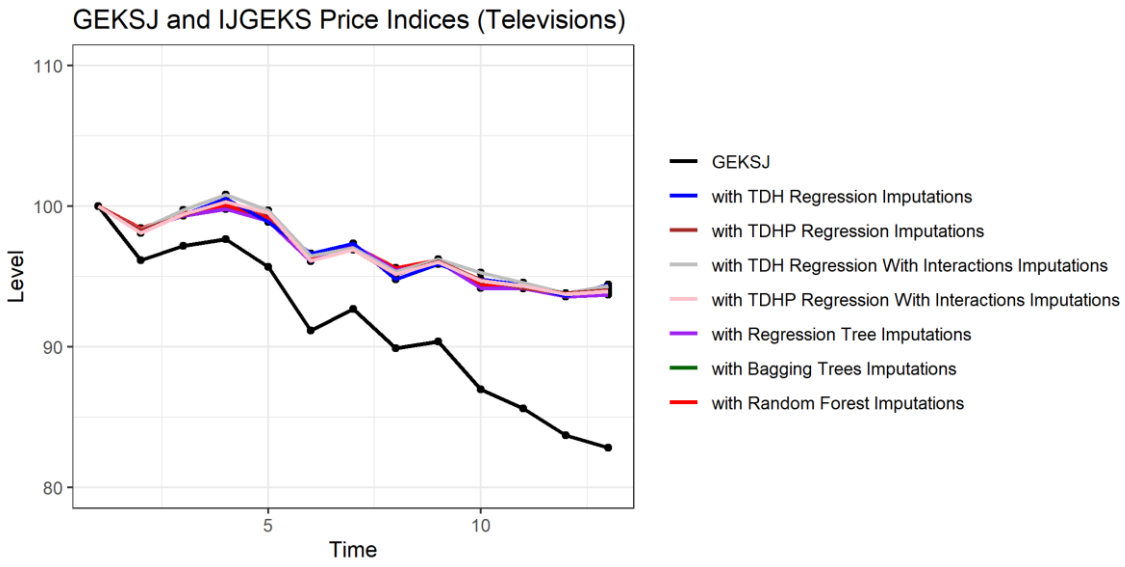


5.2 Empirical results for IJGEKS method

In the case of product categories with high churn and lifecycle pricing, the IJGEKS method can be used to compile individual price indices. This is the case of consumer electronics, in particular tablets, laptops, televisions and smartphones, which make up approximately 1 % of the CPI basket. We compiled GEKS-Jevons and IJGEKS price indices for the first three product categories by using full monthly data (all weeks are taken into account). For the IJKEKS index, the imputation methods listed in Table 2 were used. Before computing price indices, data cleaning and variables selection was performed. The variables selection was conducted by using variance inflation factors and lasso regression techniques. These techniques were implemented by using glmnet package in R. A detailed description of data cleaning and variables selection procedures are not provided here. Note that individual products are equally weighted as detailed weighting information is not available from online data. The window length for these multilateral price indices is 13 months.

Figure 7: GEKS-Jevons and IJGEKS indices (full window of 13 months)





As can be seen in Figure 7, a clear downward drift can be seen in GEKS-Jevons indices for laptops and televisions, while it is less pronounced for tablets. This is due to the systematic price downward trend of many products during their lifecycle in these product categories (see examples in Figure 4 for televisions), which does not reflect a real, constant quality, aggregate price evolution. The bilateral imputation Jevons indices include price comparisons of unmatched products. In particular, discounted prices are compared with imputed "normal" prices before leaving the assortment. In this way, the IJGEKS method allows to tackle this downward drift. The IJGEKS indices in Figure 7 suggest that this method allows correcting for the downward drift observed in GEKS-Jevons indices.

Importantly, the results of Figure 7 do not provide a clear response to the question of which price imputation method should be preferred. In order to judge about the performance[8] of the different price imputation methods, the coefficients of determination (R^2 values), which show how well a price imputation method explains variability in prices, are provided in Table 3. Moreover, comparisons of observed and of predicted (imputed) prices, which illustrate how well a price imputation method performs in terms of model fit, are provided in Figures 8, 9 and 10.

Table 3: R^2 values of the Price Imputation Methods (average for 20 samples)

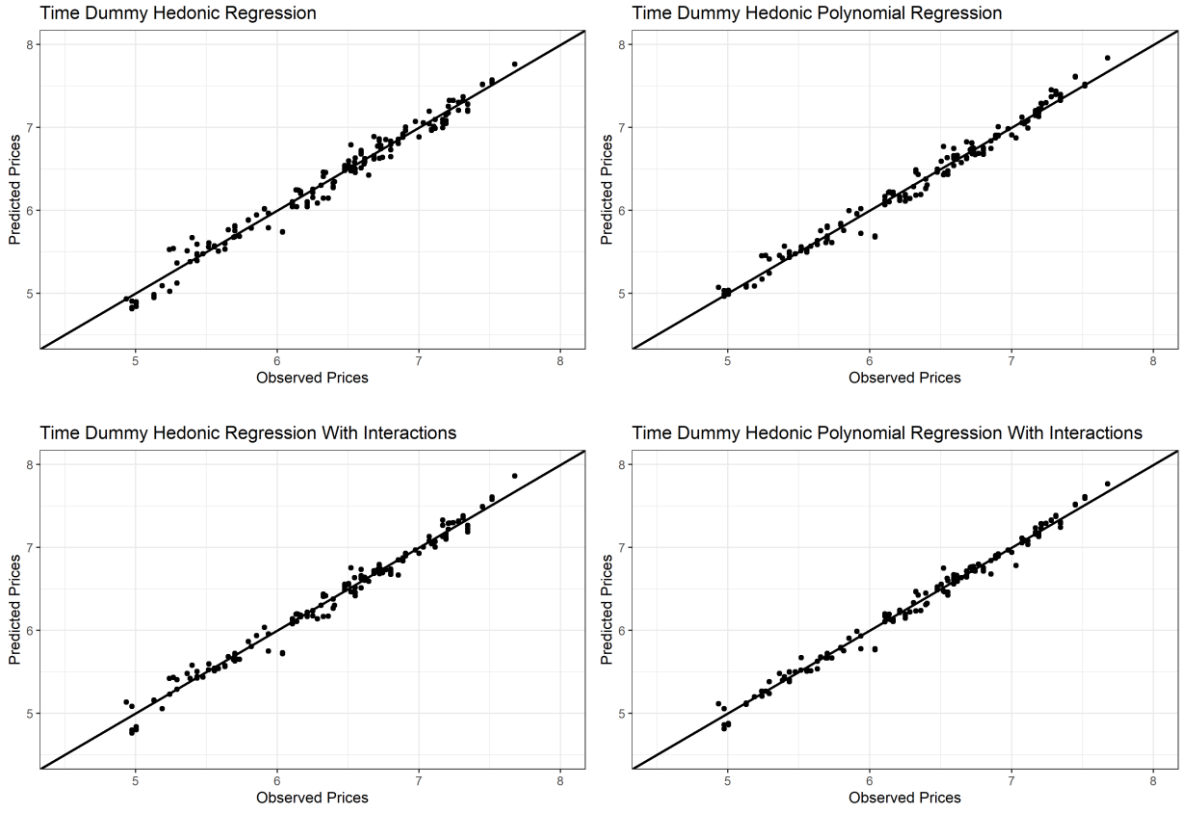
	Tablets	Laptops	Televisions
Time Dummy Hedonic Regression	0.9727873	0.8898543	0.8913327

[8] Both performance metrics were implemented by splitting the entire dataset into training (80%) and test (20%) datasets. More specifically, observed (test) prices were used along with the prices generated by the price imputation methods derived from the training dataset to obtain R^2 values in Table 3 and observed (test) prices were compared to the prices generated by the price imputation methods in Figures 8, 9 and 10.

Time Dummy Hedonic Polynomial Regression	0.9820377	0.9004979	0.9254533
Time Dummy Hedonic Regression with Interactions	0.9828617	0.9208489	0.9155837
Time Dummy Hedonic Polynomial Regression with Interactions	0.9883839	0.9184804	0.9342877
Regression Tree	0.9879692	0.9528261	0.9693565
Bagging Trees	0.9902542	0.9658194	0.9765517
Random Forest	0.9950604	0.9769152	0.9832115

Figure 8: Observed and Predicted Prices of "Tablets" Group

(for 1 out of the 20 samples)



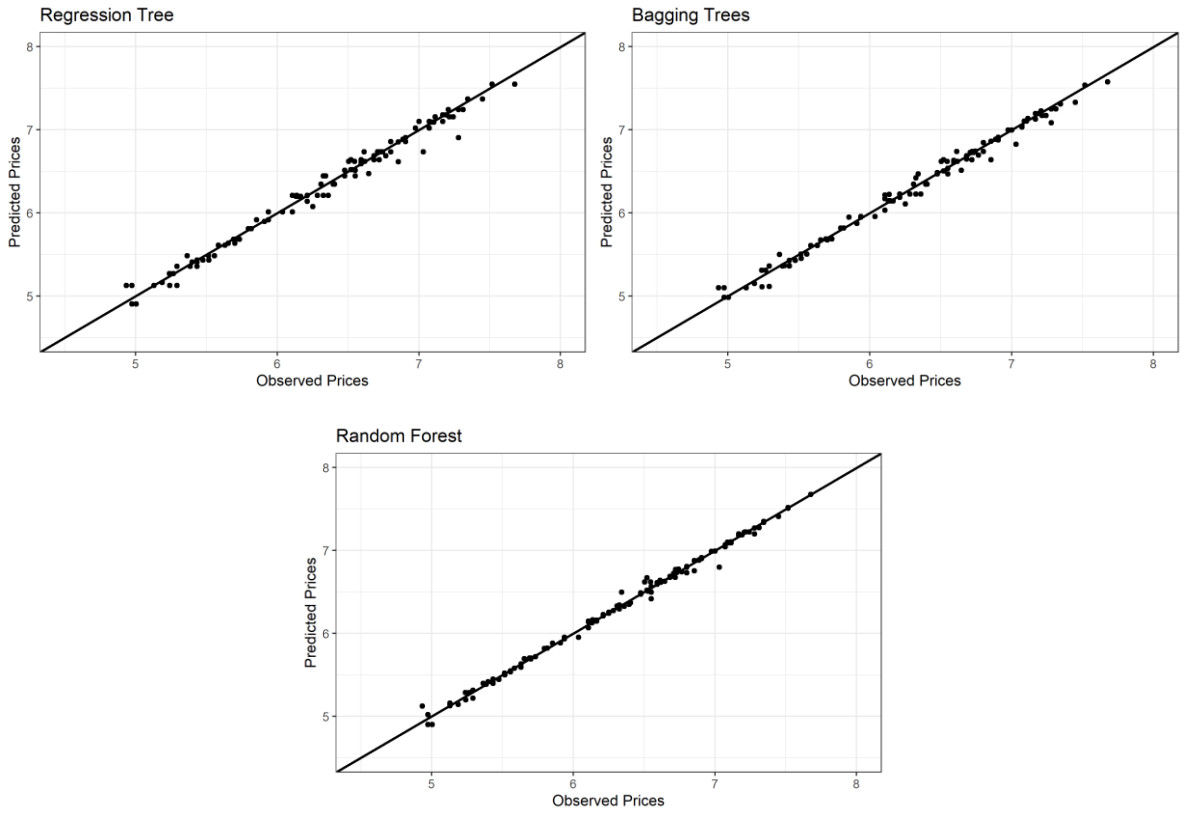
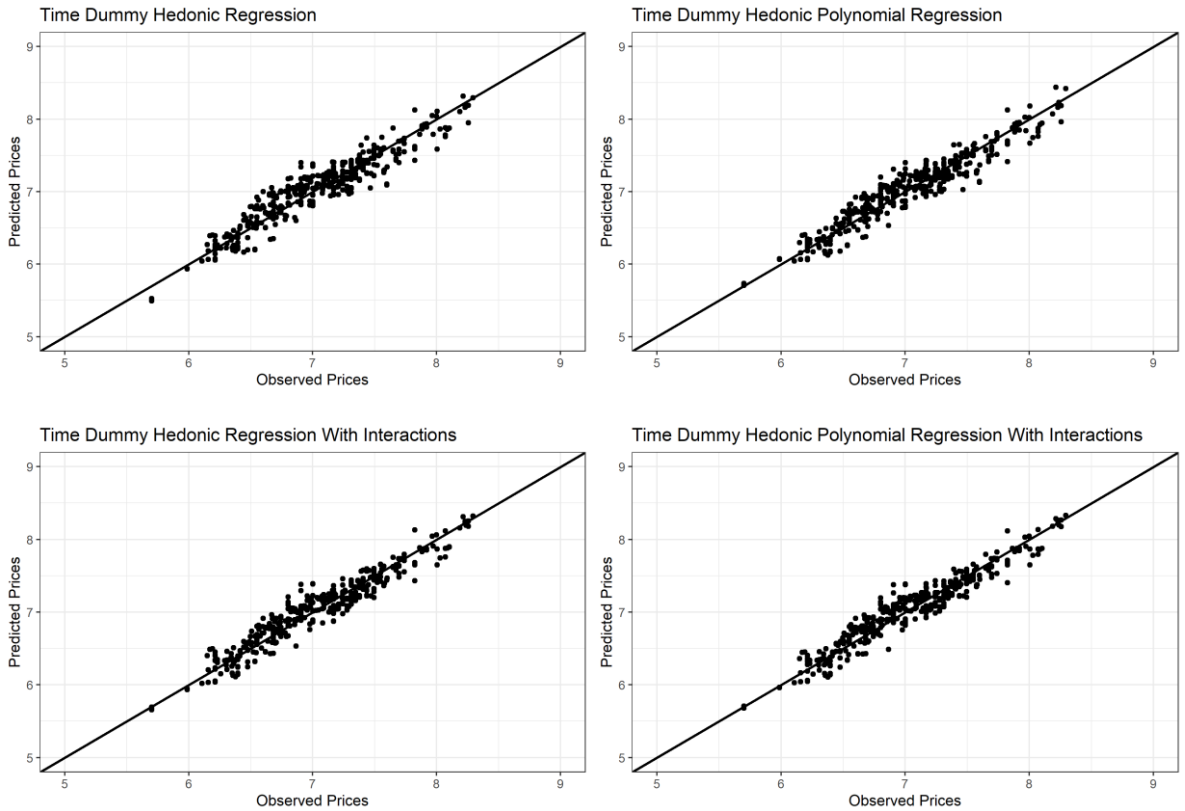


Figure 9: Observed and Predicted Prices of "Laptops" Group

(for 1 out of the 20 samples)



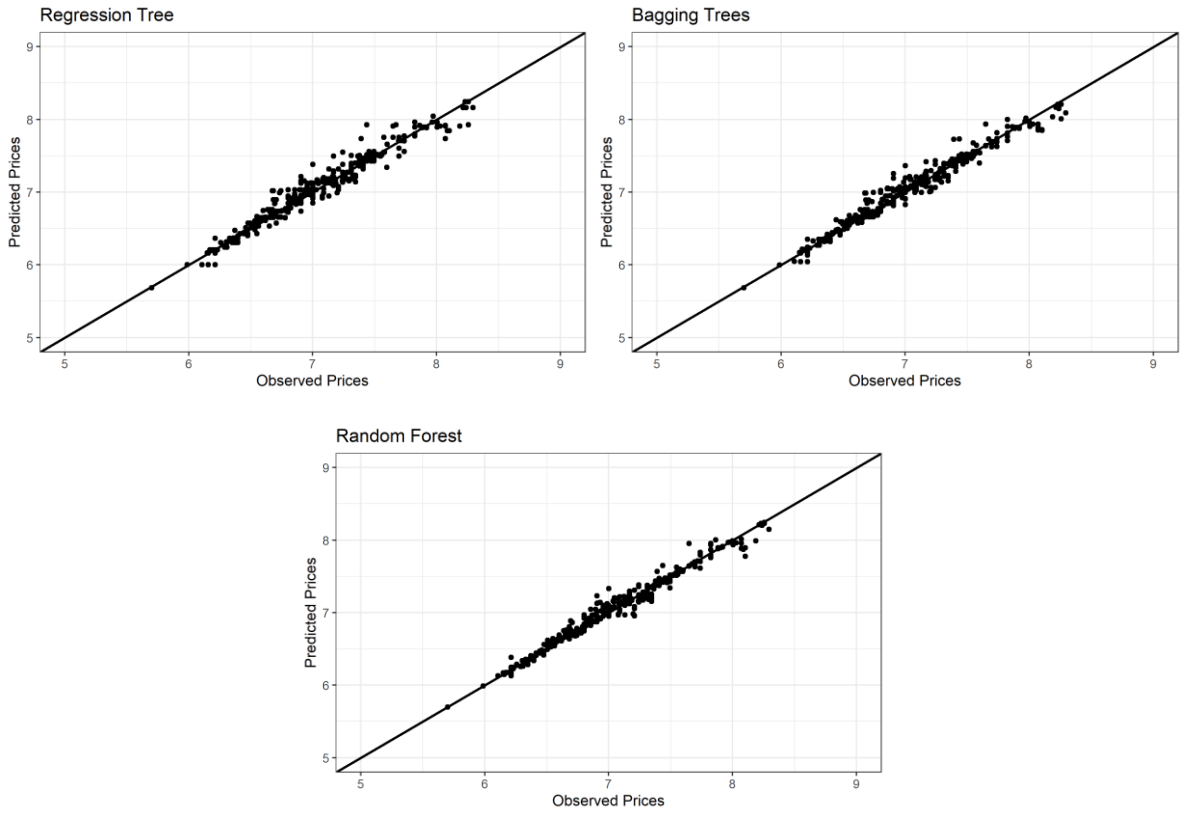
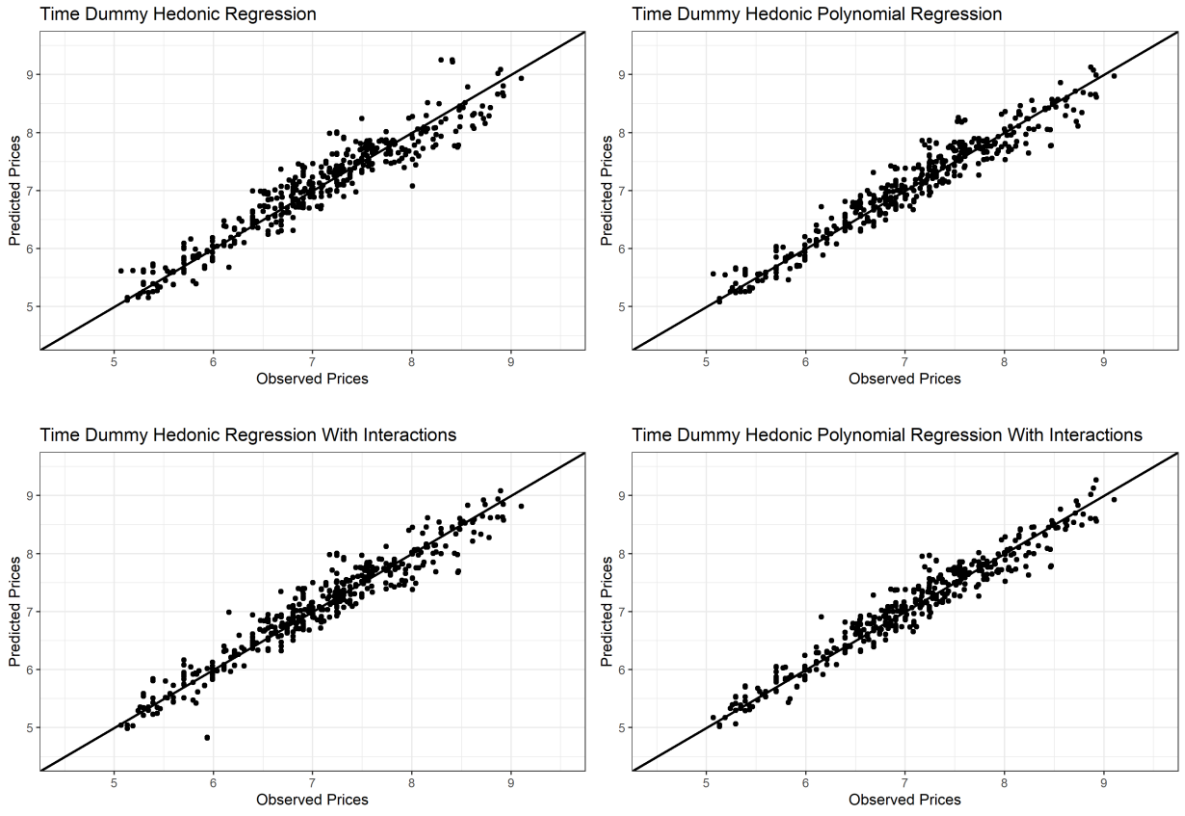
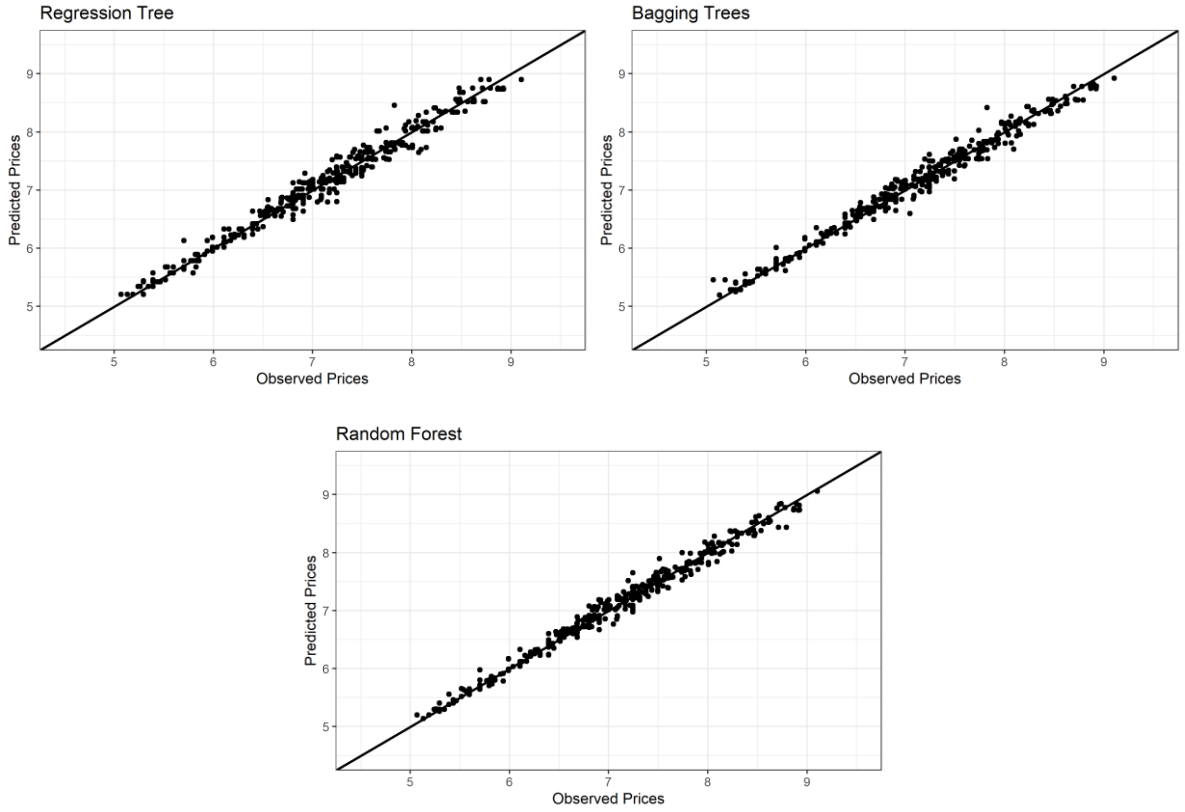


Figure 10: Observed and Predicted Prices of "Televisions" Group

(for 1 out of the 20 samples)





As it can be seen from the results of Table 3 and of Figures 8, 9, and 10, Random Forest based price imputation method provides the most accurate results both in terms of explainability of variability in prices and of model fit. Therefore, Random Forest based price imputation method might be preferred for IJGEKS method`s calculations in practice. However, before practical implementation of any price imputation method, the complexity of the method`s implementation should also be considered.

6. Conclusion

STATEC has been bulk scraping prices and product characteristics from a major website for household appliances and consumer electronics for almost two years now. Online and in-shop sales are covered at the same time with this data source as both prices and the scope of products sold are known to be the same for this retailer. The aim of this paper was to search for practical solutions to exploit the full potential of this data. Our analysis has shown that, depending on the churn of the assortment and the presence of lifecycle pricing, different methods need to be used for different product categories.

For product categories with relatively stable assortments, namely small household appliances, electrical appliances for personal care and some other product categories like printers (20 product categories in total), the multilateral GEKS-Jevons method seems to be suitable. Only products in stock are taken into account in this case. We expect to introduce this method for this data source in 2024. Questions related to the choice of splicing method as well as to the window length itself are not addressed in this paper.

High churn product categories with presence of lifecycle pricing are highly challenging. We showed that an imputation Jevons GEKS method allows in principle to address the lifecycle problem and quality adjustment issues. Different imputation

methods, hedonic or machine learning based, give very similar results. One of the main issues is the absence of individual product weights as it is not clear to what extend they would influence the result. Moreover, these methods are quite resource demanding (data cleaning, model update, analysis of the plausibility of the results, etc.) and more experience needs to be collected to judge about the feasibility of introducing these methods in production.

Other categories with relatively high churn, namely big household appliances (freezers, refrigerators, dishwashers, etc.) have not been analyzed in detail yet. One needs to check if GEKS-Jevons indices are biased due to the absence of explicit quality adjustments, which can be done by comparing them with imputation Jevons GEKS indices for example. This will be carried out in future work.

Currently, we are setting up another web-scraping project for another major retailer of consumer electronics and household appliances and the approach and lessons learned in this project are highly relevant and valuable.

List of References:

Auer, J., and Boettcher, I. (2017). From price collection to price data analytics. Statistics Austria.

Blaudow, C., and Burg, F. (2018). Dynamische Preissetzung als Herausforderung für die Verbraucherpreisstatistik. WISTA, 2(2018), 11-22.

Blaudow, C., and Seeger, D. (2019). Fortschritte beim Einsatz von Web Scraping in der amtlichen Verbraucherpreisstatistik–ein Werkstattbericht. WISTA–Wirtschaft und Statistik (No. 4, pp. 19-30).

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. Routledge.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

de Haan, J. (2010). Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and 'Re-Pricing' Methods. Jahrbücher für Nationalökonomie und Statistik, 230(6), 772-791.

de Haan, J., and Hendriks, R. (2013). Online data, fixed effects and the construction of high-frequency price indexes. In Economic Measurement Group Workshop (pp. 28-29).

de Haan, J., and Diewert, W. E. (2017). Quality change, hedonic regression and price index construction. In 15th meeting of the Ottawa Group, Eltville, Germany.

Eltetö, O. and P. Köves (1964). On a Problem of Index Number Computation Relating to International Comparisons, Statisztikai Szemle 42, 507-518 (in Hungarian).

EUROSTAT (2020). Practical guidelines on web scrapping for the HICP.

Gini, C. (1931). On the circular test of index numbers. Metron, 9(9), 3-24.

Griffioen, R., and Ten Bosch, O. (2016). On the use of internet data for the Dutch CPI. In UNECE Meeting of the Group of Experts on Consumer Price Indices, Geneva.

Guerreiro, V., Walzer, M., and Lamboray, C. (2018). The use of supermarket scanner data in the Luxembourg consumer price index. STATEC.

Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Konny Crystal G., Williams Brendan K. and Friedman David M. (2022). Big Data in the US Consumer. Big Data for Twenty-First-Century Economic Statistics, 79, 69.

Radjabov, B., and Ferring, M. (2021). The Implementation of a Multilateral Price Index Method for Scanner Data in the Luxembourg CPI. STATEC.

Szulc, B. (1964). Indices for Multiregional Comparisons, Przeglad Statystycny 3, 239-254.

van Loon, K., and Roels, D. (2018). Integrating big data in the Belgian CPI. UNECE Meeting of the Group of Experts on Consumer Price Indices.

von Auer, L., and Brennan, J. E. (2007). Bias and inefficiency in quality-adjusted hedonic regression analysis. Applied Economics, 39(1), 95-107.

von Auer, L. (2007). Hedonic price measurement: the CCC approach. Empirical Economics, 33(2), 289-311.

Annex A

Provisional list of product categories for which we intend to introduce GEKS-Jevons indices in 2024 for our new data source.

COICOP code	COICOP name
05.3.1.3.1.1	Electrical oven
05.3.1.3.1.2	Microwave oven
05.3.1.3.1.3	Induction hotplates
05.3.1.4.1.1	Hoot
05.3.1.4.1.2	Dehumidifier
05.3.1.5.1.1	Vacuum cleaner
05.3.2.1.1.1	Kitchen robot
05.3.2.2.1.1	Coffee-machine
05.3.2.2.1.2	Electric kettle
05.3.2.3.1.1	Iron
05.3.2.4.1.1	Toaster
05.3.2.9.1.1	Deep fryer
08.2.0.1.1.1	Fixed telephone equipment
09.1.1.1.1.1	Sound recording and reproduction equipment
09.1.1.2.1.1	Sound and image reproduction equipment (excluding TVs)
09.1.3.2.1.1	Printer and scanner
09.1.4.9.1.1	Memory cards
09.1.4.9.1.2	USB sticks
12.1.2.1.1.1	Hair care (hair dryer, etc)
12.1.2.1.1.2	Electric toothbrush
12.1.2.1.1.3	Electrical rasor

Annex B

This appendix gives supplementary information on the different price imputation methods, which have been tested.

The Chosen Price Imputation Methods

Hedonic Linear Regression Based Price Imputation Methods

Time Dummy Hedonic Regression

One of the most common variants of hedonic linear regression used in price statistics, when products` quality characteristics are available, is time dummy hedonic regression (de Haan 2010, pp. 779-781). Time dummy hedonic regression, in which natural logarithms of prices are modelled using time and products` quality characteristics, can be defined as:

$$\ln p_i^t = \alpha + \sum_{\tau=1} \delta^\tau D^\tau + \sum_{k=1} \beta_k z_{jk} + \varepsilon_i^t \qquad (9)$$

In Equation 9, $\ln p_i^t$ denotes natural logarithm of price of a product i in a month t and D^τ denotes time dummy variable which has a value of 1, if observation relates to a month τ, and which has a value of 0 otherwise. Moreover, z_{jk} denotes the kth products` quality characteristic variable of a product j. $\delta^\tau$ and $\beta_k$ coefficients denote time dummy and the kth products` quality characteristic coefficients. Additionally, α denotes intercept and ε_i^t denotes normally distributed error with variance σ^2 - $\varepsilon_i^t \sim N(0, \sigma^2)$. Conventionally, an estimated time dummy hedonic price index, $P_{TDH}^{0,t}$, is equal to $\exp(\hat{\delta}^t)$.

Time Dummy Hedonic Polynomial Regression

To account for nonlinearity in products` quality characteristics, time dummy hedonic regression defined in Equation 9 can be modified to include quadratic[9] products` quality characteristics[10]. Therefore, time dummy hedonic polynomial regression can be defined as:

$$\ln p_i^t = \alpha + \sum_{\tau=1} \delta^\tau D^\tau + \sum_{k=1} \beta_k z_{jk} + \sum_{k=1} \gamma_k z_{jk}^2 + \varepsilon_i^t \qquad (10)$$

In Equation 10, in addition to D^τ and to z_{jk} variables, z_{jk}^2 denotes the quadratic kth products` quality characteristic variable of a product j. Moreover, in addition to $\delta^\tau$ and to $\beta_k$ coefficients, $\gamma_k$ denotes the quadratic kth products` quality characteristic coefficient.

[9] In principle, an inclusion of polynomials with higher degrees is also possible. However, such an inclusion results in an increasing risk of overfitting.

[10] As squaring of time and of products` quality characteristics, which are expressed as dummy variables, is not meaningful, only squared products` quality characteristics, which are continuous, are taken into account while defining time dummy hedonic polynomial regression and time dummy hedonic polynomial regression with interactions in this research.

Time Dummy Hedonic Regression with Interactions

To account for interconnections between products` quality characteristics, time dummy hedonic regression defined in Equation 9 can be modified to include interactions between products` quality characteristics[11]. Therefore, time dummy hedonic regression with interactions can be defined as:

$$\ln p_i^t = \alpha + \sum_{\tau=1} \delta^\tau D^\tau + \sum_{k=1} \beta_k z_{jk} + \sum_{l=1} \sum_{l\neq k=1} \theta_{lk} z_{jl} z_{jk} + \varepsilon_i^t \qquad (11)$$

In Equation 11, in addition to D^τ and to z_{jk} variables, $z_{jl}z_{jk}$ denotes an interaction variable between the lth and the kth products` quality characteristics of a product j, where $l \neq k$. Moreover, in addition to $\delta^\tau$ and to $\beta_k$ coefficients, $\theta_{lk}$ denotes an interaction coefficient between the lth and the kth products` quality characteristics, where $l \neq k$.

Time Dummy Hedonic Polynomial Regression with Interactions

To account for nonlinearity in products` quality characteristics and for interconnections between products` quality characteristics at the same time, time dummy hedonic regression defined in Equation 9 can be modified to include both quadratic products` quality characteristics and interactions between products` quality characteristics. Therefore, time dummy hedonic polynomial regression with interactions can be defined as:

$$\ln p_i^t = \alpha + \sum_{\tau=1} \delta^\tau D^\tau + \sum_{k=1} \beta_k z_{jk} + \sum_{k=1} \gamma_k z_{jk}^2 + \sum_{l=1} \sum_{l\neq k=1} \theta_{lk} z_{jl} z_{jk} + \varepsilon_i^t \qquad (12)$$

An advantage of the above mentioned price imputation methods lies in a possibility to relatively easily explain coefficients obtained from these methods[12]. A disadvantage of the above mentioned price imputation methods lies in the fact that products` quality characteristics coefficients of these methods are time invariant. This disadvantage can be tackled by making products` quality characteristics coefficients of these methods time variant (von Auer 2007, pp. 292-293) (von Auer and Brennan 2007, pp. 96-101). Another disadvantage of the above mentioned price imputation methods lies in a global nature of these methods, which are designed to create a single predictive model for an entire data space. Importantly, if relationships between products` quality characteristics are complex and non-linear, a creation of a single predictive model for an entire data space, which fully takes these relationships into account, is very cumbersome. Therefore, a strategy to split an entire data space into smaller distinct and non-overlapping regions, where complex and non-linear relationships between products` quality characteristics are more

[11] An aim of hedonic linear regression is to model a relationship between natural logarithms of prices and products` quality characteristics during some time and not to model a relationship between natural logarithms of prices and interactions of products` quality characteristics with time. Therefore, only interactions between products` quality characteristics are taken into account while defining time dummy hedonic regression with interactions and time dummy hedonic polynomial regression with interactions in this research.

[12] With increasing complexity of hedonic linear regression based price imputation methods, explainability of coefficients obtained from these methods becomes more and more difficult.

manageable to be modelled, seems to be a viable strategy. This strategy is a foundation of tree based price imputation methods.

Tree Based Price Imputation Methods

To describe tree based price imputation methods[13], it is useful to define a general form of hedonic linear regression first. Therefore, a general form of hedonic linear regression can be defined as:

$$\ln p_i^t = h\left(D^\tau, z_{jk}\right) + \varepsilon_i^t \qquad (13)$$

In Equation 13, $h\left(D^\tau, z_{jk}\right)$ denotes function of time and of products` quality characteristics.

Regression Tree

Unlike hedonic linear regression based price imputation methods, regression tree (Breiman et al. 1984, pp. 18-58) models natural logarithms of prices using time and products` quality characteristics in a non-parametric way. More specifically, regression tree splits all observations into distinct and non-overlapping regions, $R_1 \dots R_M$, such that for any region, R_m, an average value of natural logarithms of prices of that region, $\overline{\ln p}_{R_m}$, solves a minimization problem of $h\left(D^\tau, z_{jk}\right)$ estimation, which can be defined as:

$$\overline{\ln p}_{R_m} = \operatorname{argmin}_{\hat{h}\left(D^\tau, z_{jk}\right)} \left(\sum_{\ln p_i^t \in R_m} \left(\ln p_i^t - \hat{h}\left(D^\tau, z_{jk}\right) \right)^2 \right) \qquad (14)$$

In Equation 14, $\hat{h}\left(D^\tau, z_{jk}\right)$ denotes a prediction of $h\left(D^\tau, z_{jk}\right)$. Importantly, $\overline{\ln p}_{R_m}$ is used as a regression tree prediction for natural logarithms of prices, which fall in R_m. An aim of regression tree is to find a set of distinct and non-overlapping regions such that this set solves a minimization problem, which can be defined as:

$$\{R_1 \dots R_M\} = \operatorname{argmin}_{\{R_1 \dots R_M\}} \left(\sum_{R_m} \sum_{\ln p_i^t \in R_m} \left(\ln p_i^t - \overline{\ln p}_{R_m} \right)^2 \right) \qquad (15)$$

As finding an optimal set of distinct and non-overlapping regions might be computationally infeasible for a large number of observations, regression tree uses a recursive binary splitting approach, which is described below. For any time or for any products` quality characteristic variable, x_k, regression tree selects a splitting point, ξ_k, to split an entire data space into two parts such that $\Xi_L = \{\ln p_i^t \mid x_{ik} < \xi_k\}$ denotes the left part and that $\Xi_R = \{\ln p_i^t \mid x_{ik} \geq \xi_k\}$ denotes the right part. Importantly, an optimal ξ_k value is selected in a way to solve a mean squared error (MSE) minimization problem, which can be defined as:

$$\mathrm{MSE(RT)} = \min_{\xi_k} \left(\frac{1}{N_L} \sum_{\ln p_i^t \in \Xi_L} \left(\ln p_i^t - \overline{\ln p_i^t}_{\Xi_L} \right)^2 + \frac{1}{N_R} \sum_{\ln p_i^t \in \Xi_R} \left(\ln p_i^t - \overline{\ln p_i^t}_{\Xi_R} \right)^2 \right) \qquad (16)$$

[13] Tree based price imputation methods were implemented by using caret package in R.

In Equation 16, N_L and N_R denote numbers of observations of the left part and of the right part. Moreover, $\overline{\ln p_{i\Xi_L}^t}$ and $\overline{\ln p_{i\Xi_R}^t}$ denote average values of natural logarithms of prices of the left part and of the right part. At a root node (at the first splitting) of regression tree, MSE(RT) is computed for every time and for every product's quality characteristic variable. An optimal split is selected by selecting the time or the products` quality characteristic variable and the ξ_k value, which result in the lowest MSE(RT). After this, splitting is repeated at internal nodes. Importantly, a recursive binary splitting approach is a greedy approach as it makes an optimal split at each node without accepting a possibility that a better split might be available if several nodes ahead are analyzed.

Bagging Trees

To improve predictions` accuracy achieved by regression tree, Breiman (1996, pp. 132-137) proposes to use a bagging approach, which is also called a bootstrap aggregation approach. Unlike obtaining one value for each regression tree prediction, a bagging approach proposes to use an average value of all predictions, which are obtained from multiple samples drawn from all observations with replacement. This implies that, if a dataset of N observations is denoted as $\Phi = \{(\ln p_i^t, D^\tau, z_{jk}) \mid j = 1, \dots, N\}$, a bagging approach draws M (M<N) random observations from Φ with replacement B times to create B bootstrap samples, $\Phi_b, b = 1, \dots, B$. As each bootstrap sample can be used to create a model to obtain predictions, an average value of all predictions that are obtained from all models created from all bootstrap samples is a bagging approach prediction. When a bagging approach is used on regression trees, bagging trees are created with regression trees grown on all bootstrap samples. Importantly, an average value of predictions from all regression trees is used as a bagging trees prediction for natural logarithms of prices, which can be defined as:

$$\hat{H}(D^\tau, z_{jk}) = \frac{1}{B}\sum_b \hat{h}_b(D^\tau, z_{jk}) \qquad (17)$$

Random Forest

Even if a bagging approach improves predictions` accuracy achieved by regression tree, this approach has a disadvantage. More specifically, improvements in predictions` accuracy achieved by bagging trees might not be very pronounced if predictions of these trees are correlated. To reduce correlation between predictions of bagging trees, Ho (1995, pp. 280-282) proposes a stochastic approach of selecting a random subset of independent variables when growing each bagging tree. Breiman (2001, pp. 10-11) extends Ho`s stochastic approach by allowing for a selection of a random subset of independent variables at each node when growing each bagging tree. This is a general approach of how to grow random forest. This implies that, if an entire data space of time and of products` quality characteristics variables is denoted as $X = \{x_1, \dots, x_K\}$, a subset of time and of products` quality characteristics variables, X_S (X_S< X), is drawn from X at each node for each bagging tree to obtain a collection of trees, which is called random forest. Importantly, an average value of predictions from all trees from such a collection of trees is used as a random forest prediction for natural logarithms of prices.