ESSnet Trusted Smart Statistics – Web Intelligence Network

Grant Agreement Number: 101035829 — 2020-PL-SmartStat

Work Package 3

New Use-cases

Deliverable 3.7:
UC3: Report on methodology and results for online prices

Version, 2025-02-24

Prepared by:

UC coordinator(s):

Petrus Munter – Statistics Sweden (SCB); Petrus.munter@scb.se

Contributors:

Petrus Munter – (SCB, Sweden)

Remy Kamali – (SCB, Sweden)

Can Tongur – (SCB, Sweden)

Jens Andersson – (SCB, Sweden)

Kostadin Georgiev – (BNSI, Bulgaria)

Olav ten Bosch – (CBS, the Netherlands)

This document was funded by the European Union.

Inhoudsopgave

1 Background

This document is part of the Work package 3 (WP3) *New use-cases* from the ESSnet Trusted Smart Statistics – Web Intelligence Network project (ESSnet TSS-WIN). The overall objective of WP3 is to explore the potential of new types of web data sources for official statistics, with each use-case focused on a specific use case (UC). The set of use cases being explored are:

- **UC1** Characteristics of the real estate market
- **UC2** Construction activities
- **UC3** Online prices of household appliances and audio-visual, photographic and information processing equipment (and generalising the data-collection to other activities)
- **UC4** Experimental indices in tourism statistics (hotel prices)
- **UC5** Business register quality enhancement
- **UC6** Faster Economic Indicators using new data sources

This particular deliverable focusses on the most interesting outcomes of the work done in UC3 on online prices on audio-visual, photographic and information processing equipment. In previous years, as reported in year reports D3.1, D3.2 and D3.3 price collection and experimentation has been done in two countries: Sweden and Bulgaria. The work consisted of exploration of the data sources, development of web scraping software, data acquisition and recording and processing of the data. This has led to the conclusion that the execution of these process steps on multiple data sources in different countries is very well possible, but also that from a statistical viewpoint it could be more valuable to focus on opportunities for finding and estimation distributions regarding the correlation between popularity of each item and how much they actually sell by combining web data with administrative data.

Therefore, in this document we only briefly highlight the early work of the first years in two countries in chapter 2. For more detailed information on the experiments and experiences in the early approaches we refer to the year reports D3.1, D3.2 and D3.3. Because of the positive outcomes and general applicability of the work on combining data sources the sequel of the document, from chapter 3 onwards, we focus on the pilot to combine web data with administrative data, a cash register in particular, in a very specific methodologically sound way. Statistics Sweden tested this on their data sources, which has led to the insight that it could be quite an interesting approach to enhance consumer price index measurements through data integration. It was tested in a proof of concept which has been described so that it can be repeated by other statistical organisations on their specific data landscapes.

2 Processing steps 1 to 4 in Sweden and Bulgaria

In the first years of the project the focus was on executing process steps 1 to 4 on the data sources of interest in this use case (1) New data sources exploration (2) Programming, production of software (3) Data aquisition and recording and (4) Data processing. The goal was collection of data about online prices of household appliances and audio-visual, photographic and information processing equipment by web scraping of online shops. The outcomes have been described in year reports D3.1, D3.2 and D3.3, available on the wiki of the WIN project, from which we here give a very brief summary.

Statistics Sweden pulled data from 2 online sources with several scraping sessions stretched throughout each week to get a good and accurate estimate of the online market. The cash register data they received through deliveries from each company was combined with the web-scraped data, creating a dataset that allows to link articles to specific product groups, companies, and sales weeks. This combined dataset also made it possible to calculate average prices from each online source and from the cash register data.

One of the challenges was discontinued software support for web scraping software. This was solved by changing to Python for web scraping. Another challenge came up when one of the companies collecting data from got bought and changed its name at the end of the year. This caused all the article numbers to change, making it almost impossible to maintain a continuous timeline. Such ownership changes were a new phenomenon for statistics Sweden. It makes it necessary to match products over time based on other parameters than the article number, for instance by using product names. Other challenges were changes on the website, software changes making code outdated/invalid and the dependency on specific skills that are usually centred around just a few people within the agency. These were however manageable.

Statistics Bulgaria Started initially with four online data sources but after a re-evaluation using the checklist developed earlier, three remained. The products scraped are products such as washing machine, refrigerator, electric steam iron, blender, espresso coffee machine and TV. Regular collection started April 2022 proceeding every Friday of every week. It was performed early in the morning to reduce burden on sites and with a moderate delay between requests. During the first project year, the software for collection and processing was developed, tested and implemented. During the second and the third project year only minor adjustments and updates were necessary. The extracted text blocks were stored in fields in csv files to be processed right after data collection via dedicated processing scripts. These scripts were unified for data sources and sent notifications to experts after finishing. When changes in the site structure or the description of the products occurred, the processing scripts were adjusted and another iteration was performed. The variable “price” was converted from BGN to Euro and descriptive statistics - min, max, mean, standard deviation for each data source for all products of interest were calculated. Moreover, for the purposes of production of experimental statistics for the prices of the household appliances of interest, the following indicators were defined, after consulting price statisticians, in order for the outputs to be viable for use in official CPI calculation at a later stage:

- Monthly average price per product item, individually for every single online shop;
- Monthly average price per product item, aggregated for all on-line shops;
- Monthly “elementary” price index - unweighted per product item for all on-line shops (base period – April 2022) calculated as follows: (Average price month n 2022/Average price April 2022) *100 where n is the month of interest;
- Chain monthly price index per product item - measures changes in average prices, (Average price month n/ Average price month n-1) *100 where n is the month of interest

Statistics Bulgaria identified several challenges applying not only to the area of household appliances but also to clothing online shops being explored for their potential. The most prominent challenge remains the lack of scanner data available for BNSI. This leads to the inability to observe the quantities purchased online and subsequently the inability to calculate weighted indices.

Another major issue is the classification in different NACE categories of online shops and their corresponding physical stores, because the services provided are different. In addition some physical stores have no website. Therefore small regional shops with no online presence are excluded from the scope.

3 The Swedish pilot on combining data sources

Statistics Sweden has leveraged the opportunity provided by the ESSNET Project Use Case 3 to explore the potential of integrating multiple data sources to enhance our measurements for the Consumer Price Index (CPI). A continuous and central discussion within the CPI framework revolves around the handling of weights, both in terms of weighting different product groups or COICOP categories and in weighting each product within our sample.

This report serves as a proof of concept, investigating the feasibility of combining cash register data with web-scraped data to gain additional insights and improve our toolkit for managing data sources, particularly in scenarios where sales numbers per product are not readily available.

4 Current state of the CPI and Web Scraping

Currently, approximately 8% of Sweden's Consumer Price Index (CPI) data is collected through web scraping. We aim to increase this percentage over time, leveraging automation tools to minimize manual processes, thereby reducing data collection time, enhancing data quality, and increasing data volume.

A significant challenge we face with web scraping is the lack of detailed sales numbers for each product, complicating our efforts to accurately weight each product in our survey. In contrast, Statistics Sweden benefits from several sources of cash register data within our CPI sample. This data provides precise information on the quantity sold per product and product group, offering a comprehensive representation of consumer purchases.

Our project's objective is to explore the feasibility of using insights from both data sources—cash register data and web-scraped data—to estimate sales quantities (in terms of weights) weights for products in companies where such information is unknown. We aim to determine whether this approach can improve our estimated average price development compared to the current method, which evenly weights each product when only web-scraped data is available.

5 Methodology

5.1 Data gathering

Web Scraping

Our web-scraped data is collected using a combination of methods involving Python and iMacros code. Currently, Statistics Sweden is transitioning to a more Python-focused web scraping approach due to the discontinuation of support for iMacros by its supplier. The use of multiple methods has been necessitated by variations between the websites being scraped.

Both API-based web scraping and basic HTML web scraping have been implemented using basic python modules such as beautifulsoup, pandas and requests.

Over the years, Statistics Sweden has developed a comprehensive toolkit for handling both HTML parsing and the utilization of underlying APIs from webpages. When scraping large volumes of data, we prefer to use APIs when available due to their speed. However, we also employ HTML parsing options when APIs are not accessible.

There is typically a trade-off when scraping multiple products simultaneously in terms of the level of detail obtained per iteration of a scraper. In practice, this distinction manifests as follows: scraping data from a webpage displaying several products at once generally yields limited specific information per product, whereas scraping data from webpages where each page displays a single product allows for the extraction of more detailed information. This detailed information often includes attributes such as article properties, online availability, and product group connections.

Cash Register Data

Statistics Sweden receives cash register data from numerous companies on a weekly basis. In this use case, we have focused on televisions and laptops from some of Sweden's largest home electronics suppliers.

The data collected over time includes information from two companies and two different product groups (laptops and televisions). This will result in a total of four time series to compare in the results section.

In recent years, Statistics Sweden has made significant efforts to enhance the use and efficiency of cash register data. We have aimed to coordinate various surveys so that when engaging with new companies, and occasionally reevaluating older sources, we propose a data format that can sufficiently replace previous data collection methods.

A common scenario involves a combination of data collection methods, including web scraping, online forms, and physical store visits. Data collected through these methods are utilized in different surveys, such as the Consumer Price Index (CPI) and turnover in the service sector. This approach presents an opportunity for Statistics Sweden to minimize redundant work. For data providers, it often represents a significant improvement, transitioning from multiple deliverables to a single one.

We strive to automate the data delivery process as much as possible, offering several technical solutions for data receipt and collection. This flexibility is crucial, as preferred methods for setting up deliverables vary significantly among companies. Some of these options include:

- Gathering data from the data provider's suggested API solution.
- Data provider delivering data via our provided API.
- SFTP solutions, where we agree on an online folder for automatic file delivery from the data provider's software.

The initial contact and the time between our first meeting with a data supplier and receiving a first test file are usually quite short. Once we connect with individuals familiar with the data, they can typically match most of the data we request in our forms. While there is variation in the "nice to have" data, there are rarely issues with the main variables we request. We categorize variables into mandatory and optional to ensure we can produce statistics from the mandatory variables and some optional ones. If a company cannot supply the mandatory data, we do not proceed with negotiations.

The mandatory variables include:

- Article ID (article number, EAN, or both)
- Transaction date
- Sold quantity
- Organization number
- Total turnover

This information alone allows us to replace other data acquisition methods for the turnover survey in the service sector and forms the basis for the CPI. However, additional variables are needed for the CPI to better understand and track products over time. These supplementary variables include article name, external supplier, brand, and product groups.

Most data suppliers organize data within different product groups, which is extremely useful for the CPI. The structure of these groups varies significantly among suppliers, with some having many levels and others only a few. Matching these group structures with other data can be challenging, as descriptive information is often stored separately from sales data.

Once the technical solution is confirmed and a steady data flow is established, our statisticians take over to integrate the data. This process is much longer than the initial setup due to the significant variety between companies and products, making standardized controls for each product group difficult. This challenge is particularly relevant for implementing cash register data into the CPI.

In practice, this means we often need additional time after setting up the initial deliverable before we can fully implement the data source. We typically resume dialogue with the company a few months after data collection begins to address any questions and finalize the integration. Our goal is to reduce this time over time, and we are continually improving in this area.

As the number of cash register data sources has increased significantly, we have adjusted our agreements with data providers. Previously, we did not legally bind suppliers to a specific delivery method, as the agreement was mutually beneficial. However, with the increase in data volume, we recognized the potential issue if multiple data providers ceased delivering data in this manner. Consequently, the CPI has implemented changes to how data is provided to Statistics Sweden.

As of 2024, we are authorized to demand a specific data source to ensure the continuity of critical data streams for the CPI. This change aims to prevent the loss of essential data without imposing undue pressure on data providers to adopt methods incompatible with their systems and preferences.

5.2 Average price calculations

The objective of this methodology is to compute three distinct average price trajectories over time. Subsequently, these trajectories will be compared to ascertain whether the application of estimated weights enhances the precision of price development estimations. The data has undergone multiple processing stages utilizing various software tools. As previously mentioned, we employed different tools, modules, and methods for web scraping. Similarly, the data processing steps involved the use of Python for certain tasks, while the majority of the processing was conducted using SQL.

The three average prices we aim to determine are:

A. Average price($\bar{x}_{crd}$) based on actual sales figures derived from cash register data. This can be expressed mathematically as:

$$\bar{x}_{crd} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Formula 1. *Where w represents the true weekly sales quantity for product i, and x denotes the average price for the same product. The summation extends from i too n, where n is the total number of overlapping products between the web-scraped data and the cash register data.*

B. Average price($\bar{x}_{ws}$) based on equal weights from web scraped data. This can be expressed mathematically as:

$$\bar{x}_{ws} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Formula 2. *Where x is weekly average price for product i and n is the total number of overlapping products between the web scraped and cash register data.*

C. Average price($\bar{x}_{est}$) based on estimated weights obtained from combined data sources from one company projected on another company where only web scraped data is available

$$\bar{x}_{est} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Formula 3. *Where w is estimated weekly sales quantity for product i and x is the average price for the same product. Average reaching from 1 to n where n is the total number of overlapping products between the web scraped and cash register data.*

Each average is calculated for each unique combination of company and product group.

The methodology for combining data sources is straightforward. Our objective is to determine the actual number of products sold based on online information. We have web-scraped data from two companies, categorizing each product group by popularity. Our goal is to estimate the weight distribution, where the y-axis represents the true number of items sold, and the x-axis represents the position on the webpage sorted by popularity. The process is illustrated under *5.5 Sales Number Estimation.*

The matching process is conducted using article numbers obtained from both cash register data and web-scraped data. The estimated distributions are based on data from each unique combination of company and product group, resulting in four separate estimations. For instance, the estimated weights for Company A's product group G, derived from combining cash register data and web-scraped data, are then applied to estimate the weights correspondingly for Company B's product group G.

It is important to note that we have both data sources for both companies and product groups. Consequently, we can determine the "true" average price development for each of the four combinations of company and product group. It is important to note that the value presented does not represent the true average for the entire COICOP group. Instead, it pertains to a single product within the COICOP group.

5.3 Time series issues

We encountered several challenges during the data collection process, which constrained the length of the available time series. Consequently, we are presenting the price development on a weekly basis rather than a monthly one.

A significant issue arose when one of the companies we were collecting data for underwent a sale and acquisition. This transition led to a comprehensive overhaul of the company's internal article number system, rendering it impossible to match products consistently over time. This disruption effectively truncated our time series data.

Currently, Statistics Sweden lacks the capability to automate web scraping using Python. Our primary web scraping tool is based on iMacros, which necessitates several manual processes to maintain the integrity of our timelines. This reliance on manual intervention has posed challenges in ensuring the consistency and continuity of our data collection efforts.

5.4 Limitations

Due to anomalies such as arbitrary stacking of special offers on web pages and the presence of noise resulting from low sales volumes for each product, irrespective of their position in the web-scraped data, we have implemented cutoffs to achieve a more reliable estimation of the distribution between popularity and sold quantity.

The sample for these estimations is derived from the first three weeks of collected data. To ensure an accurate average popularity, data was web-scraped three times per week, which was then matched with the weekly cash register data. We focused on the 30 most popular products and those with sales numbers exceeding six units.

These limitations were established based on a visual inspection of the data to enhance the robustness of our estimations.

5.5 Sales number estimation

The plots presented in this section are artificially generated and do not represent actual data, although they are designed to approximate the general trends of the distributions. The rationale behind this approach is to preserve the integrity of the suppliers' data while also providing a clear visual illustration of the process steps within this section.

Regarding the formula used to calculate the average price over time, we proceeded as follows:

For the true average price, we relied on cash register data for all products that matched the web-scraped data. This ensures that the sample for average prices is identical for both the web-scraped data and the cash register data. Each product is then weighted by to its true sold quantity per week according to subsection 5.2.

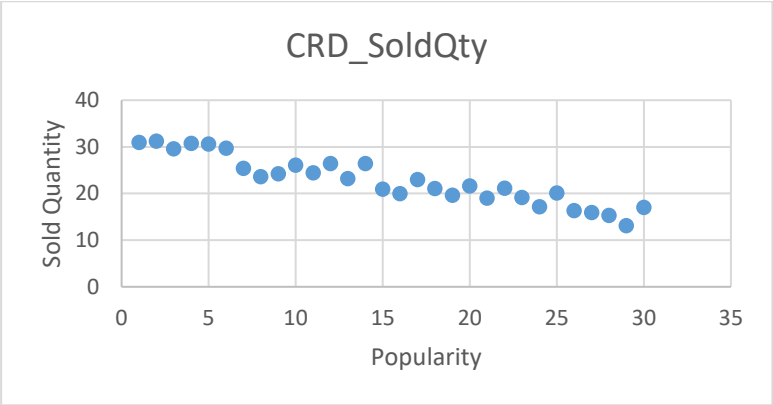

Figure 1. *Representation of downwards trending sales numbers in relation to the popularity of the product online.*

For the web-scraped data, we use the same sample as for the cash register data. However, in this scenario, we assume that the actual weights are unknown. Instead, we follow the current practice for web-scraped data by assigning equal weights to each product. From this, we derive the weekly average price for each product group and company. This approach allows us to compare the estimated average prices with those derived from the true sales quantities.

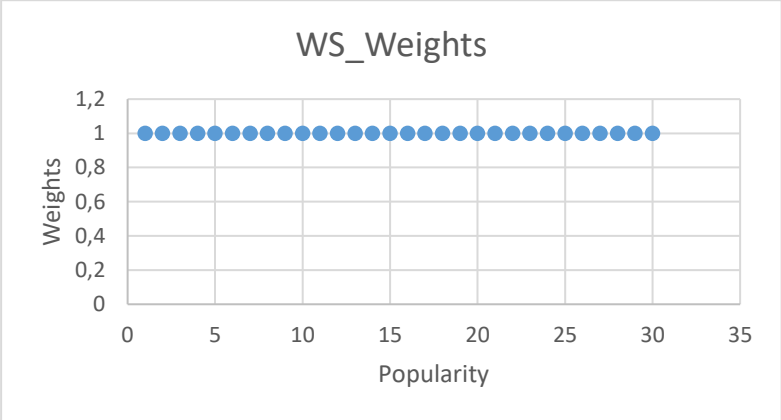

Figure 2. *Representation equal weights independent of popularity in the case of only having access to web scraped data.*

To synthesize the combined sources, we proceed as follows: We plot the data from Product Group A of Company X to examine the correlation between sales numbers and popularity for this specific product.

Subsequently, we utilize this information, or the estimated weighting function, to calculate the average price for Product Group A of Company Y.

It is crucial to highlight that we employed a simple linear regression to estimate the weights. This approach indicates that, at certain points, the estimated quantity of products sold falls below one due to the downward trend of the function. To address this, we excluded products with an estimated sold amount of less than one. The rationale behind this exclusion is straightforward: if our estimations suggest that these products are unlikely to be sold, they should not be included in the sample. Consequently, the samples used for calculating averages with this method are always a subset of those used for the web-scraped and cash register data versions.

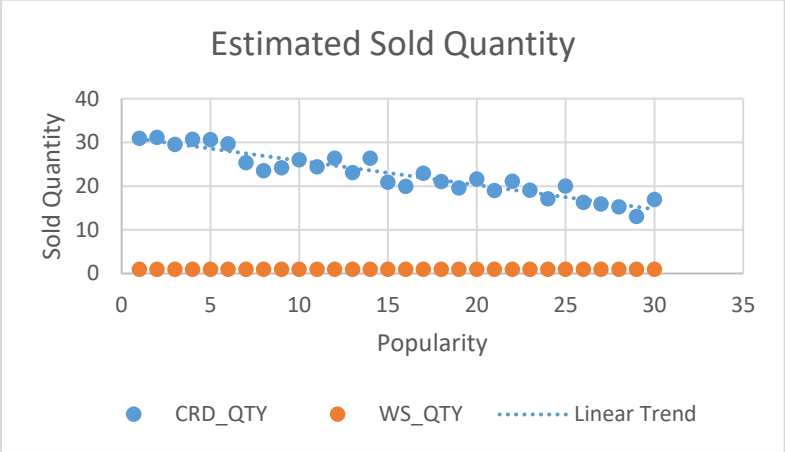

Figure 3. *Representation of the estimated sales numbers(linear trend blue dotted line) that will be used on the other company as if we didn't have sales numbers data at hand from the cash register data.*

6 Results

For the results, we present four initial graphs. These graphs illustrate the deviation from the true average price per week for each company and product group. One graph depicts the percentage deviation using the average web-scraped mean (equally weighted, represented by the orange line with squares), while the other shows the percentage deviation using the estimated sales numbers function from the other company (represented by the blue line with triangles).

The residual percentages are calculated according to:

$$RES\_PERC = \sqrt[2]{\left(\frac{y-x}{y}\right)^2}$$

Formula 4. *Where x is the average price derived from either evenly weighted web scraped data or the average derived from using the estimated weight from both sources and y is the true average derived from cash register data.*

EST_RES_PERC represents the average weekly price based on the estimated weights, expressed as a percentage deviation from the true average derived solely from the cash register data. **WS_RES_PERC** indicates the average weekly price based on web-scraped data, also expressed as a percentage deviation from the true average derived solely from the cash register data.

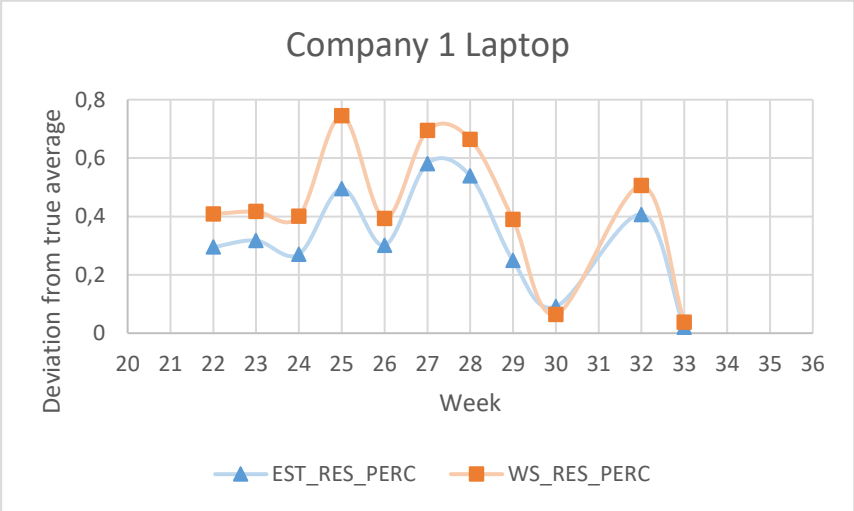

Figure 4. *Comparison of the percentage deviation from the true average derived from cash register data. Blue line with triables is percentage deviation when using estimated weights from the other company and the Orange line with squares is the percentage deviation using equal weights on the web scraped data alone. Deviation from true average reaching from 0(0%) to 0,8(80%).*

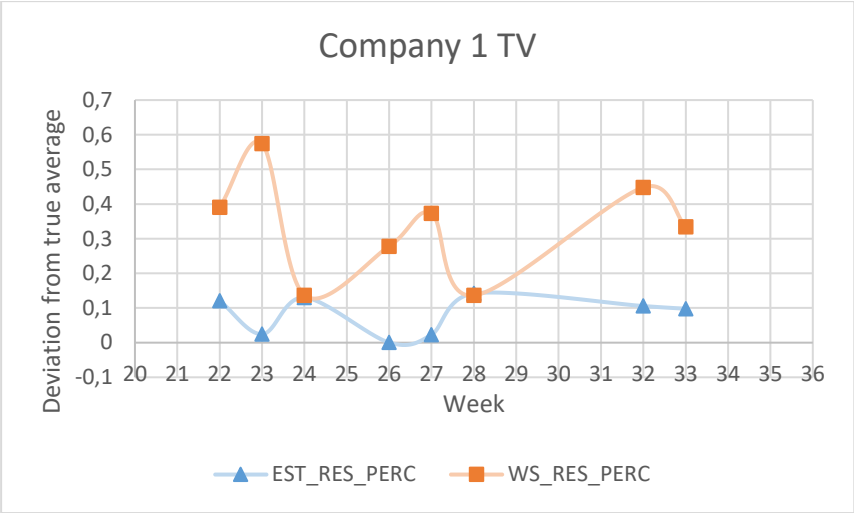

Figure 5. *Comparison of the percentage deviation from the true average derived from cash register data. Blue line with triangles is percentage deviation when using estimated weights from the other company and the orange line with squares is the percentage deviation using equal weights on the web scraped data alone. Deviation from true average reaching from 0(0%) to 0,6(60%).*

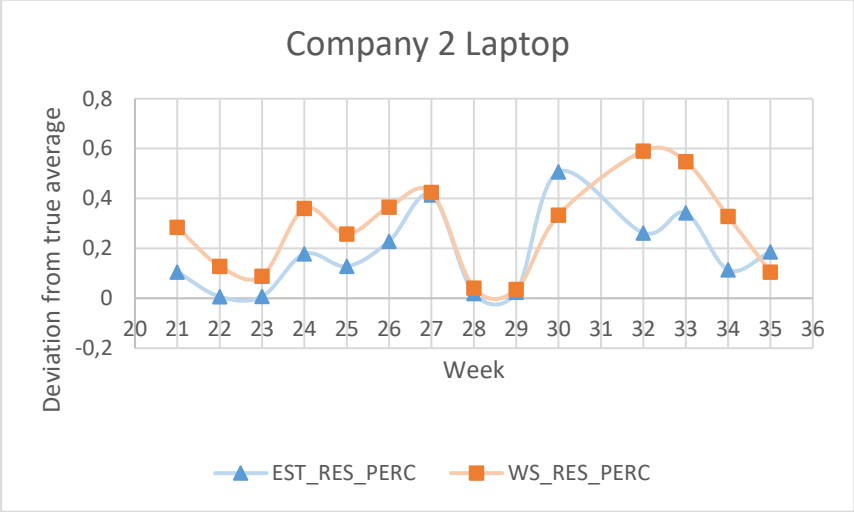

Figure 6. *Comparison of the percentage deviation from the true average derived from cash register data. Blue line with triangles is percentage deviation when using estimated weights from the other company and the orange line with squares is the percentage deviation using equal weights on the web scraped data alone. Deviation from true average reaching from 0(0%) to 0,6(60%).*

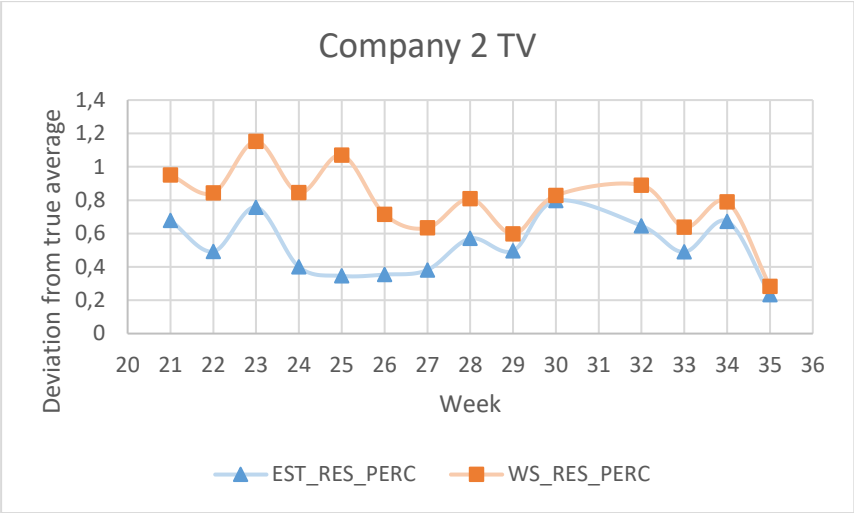

Figure 7. *Comparison of the percentage deviation from the true average derived from cash register data. Blue line with triangles is percentage deviation when using estimated weights from the other company and the orange line with orange is the percentage deviation using equal weights on the web scraped data alone. Deviation from true average reaching from 0,2(20%) to 1,2(120%).*

It is evident that the average estimations using the new method, which leverages information from the combined source of the other company, are significantly more accurate. In fact, in over 90% (91.5%) of cases across all weeks, companies, and product groups, the estimated average based on the estimated weights is closer to the true average than the estimated average prices based on web scraped data alone. It is also worth mentioning that in approximately 98% of all datapoints the average price derived from evenly weighted web scraped data was higher than the average derived when estimated weights from the combined sources were used.

7 Discussion

Statistics Sweden is pleased with these results, but we are also acutely aware of the numerous decisions made to adhere to the project's timeframes. Therefore, we will discuss several areas that might raise questions about the results and provide incentives for further exploration.

7.1 Weight estimation

We opted for a linear regression model to estimate sales numbers in relation to product popularity. The task of identifying the optimal method for these estimations could constitute a separate project. We considered whether exponential regressions might be more appropriate, potentially including less popular products at lower weights. Another approach discussed was to include only the most popular products but weigh them equally. Various methodologies could be applied to this aspect of the process, and we had to make several arbitrary decisions throughout the project. Additionally, we chose to exclude white noise from the sample used for our regression model. This decision warrants further investigation, as there might be a case for including less popular products (beyond the top 30) in the sample. This consideration also applies to products with a quantity sold below seven.

Another consideration is the initial choice to order products by popularity. Our assumption was that popularity would best describe sales numbers. However, there are other sorting options available online, such as "sort by price," "new in stock," and "special offers." Further research into these alternative sorting methods would be highly beneficial to gain a deeper understanding of the correlation between the two data sources.

7.2 Web scraping bias

We observed that the web-scraped average price is almost always higher than the average price derived from the estimated weights. This discrepancy is evident not only in the residual percentage deviation but also in the absolute average. One possible explanation is that the most expensive products tend to have lower popularity compared to more affordable options. Another consideration is the potential issues with the web scrapers. A common challenge in web scraping is the difficulty in automatically identifying what is not scraped. For instance, if a company has a fire sale and the discounted price is displayed in a different section of the webpage, there is a high chance of missing the "best" price without noticing.

It is also important to note that the scrapers were configured to scrape all products sorted by popularity. Other sorting options, such as by price or brand, are available on the websites we scraped. There might be a stronger correlation between sales numbers and other sorting criteria for displaying products online.

7.3 Product groups and companies

The sample for this project was limited to the two companies and product groups detailed in the report. This naturally raises questions about the broader applicability and reliability of this method. Expanding the research to include more product groups within the technology sector would provide a more comprehensive understanding. Additionally, exploring areas beyond technology and household items, such as clothing, could be a valuable next step.

7.4 Time

The timespan for this study is suboptimal, as the Consumer Price Index (CPI) is typically calculated on a monthly basis. Due to technical issues and time constraints, we had to compromise by examining weekly developments. Investigating the same phenomena on a monthly basis over a year could potentially yield different results.

7.5 Index Calculations

Given that the initial objective was to explore potential improvements for CPI indexes, this area certainly warrants further investigation. We regret not anticipating the challenges that necessitated a change in our analytical approach. Conducting standard index calculations, whether through a standard change process or a continuous sample, would be highly intriguing to determine how this method applies to those indexes. This could provide valuable insights and potentially enhance the accuracy and reliability of CPI measurements.

8 Final Words

This work executed in this use case consists of the systematic extraction of online price data for a set of consumer goods, i.e. household appliances and electronic equipment. This involved the identification of relevant data sources, the development and deployment of web scraping software, and the subsequent acquisition and processing of the gathered information in Sweden and Bulgaria. This showed that gathering price data from different websites in different countries on comparable phenomena is possible and that it can be maintained for multiple years.

In Sweden, the approach involved the collection of data from two online sources. Notably, the project integrated web-scraped data with cash register data, creating a dataset that allows to link articles to specific product groups, companies, and sales weeks. This combined dataset also made it possible to calculate average prices from each online source and from the cash register data. However, challenges such as discontinued software support and corporate acquisitions necessitating product re-identification were encountered and addressed, demonstrating the project's adaptability.

Statistics Bulgaria focused on the collection of online price data for various household appliances, implementing a regular data collection schedule. The project involved the development of specialized scripts for data processing and the calculation of key price indicators, including monthly average prices and price indices. However, the lack of access to scanner data presented a significant obstacle, limiting the ability to calculate weighted indices. Furthermore, discrepancies in the classification of online and physical stores, coupled with the absence of online presence for some retailers, posed challenges to ensuring comprehensive data coverage.

Inaddition to these core activities, this report describes the preliminary work done by Statistics Sweden on combining online prices with cash register data to gain additional insights and improve our toolkit for managing data sources, particularly in scenarios where sales numbers per product are not readily available. The results are promising which leads us to the conclusion that there is significant potential for further exploration in this direction. The initial goal to determine if we could extend the value of data by incorporating different sources beyond the company in question, is achieved. Statistics Sweden is eager to continue this discussion and collaborate with anyone interested in further research over the coming years.