National Institute of Economic and Social Research

OUTLIER DETECTION METHODOLOGIES FOR ALTERNATIVE DATA SOURCES: INTERNATIONAL REVIEW OF CURRENT PRACTICES

Janine Boshoff
NIESR

Xuxin Mao
NIESR

Garry Young
NIESR

NIESR Discussion Paper No. 523
02 December 2020

About the National Institute of Economic and Social Research

The National Institute of Economic and Social Research is Britain's longest established independent research institute, founded in 1938. The vision of our founders was to carry out research to improve understanding of the economic and social forces that affect people's lives, and the ways in which policy can bring about change. Over eighty years later, this remains central to NIESR's ethos. We continue to apply our expertise in both quantitative and qualitative methods and our understanding of economic and social issues to current debates and to influence policy. The Institute is independent of all party political interests.

National Institute of Economic and Social Research
2 Dean Trench St
London SW1P 3HE
T: +44 (0)20 7222 7665
E: enquiries@niesr.ac.uk
www.niesr.ac.uk
Registered charity no. 306083

This research has been funded by the Office for National Statistics as part of the research programme of the Economic Statistics Centre of Excellence (ESCoE). This paper was first published in July 2020: "Outlier detection methodologies for alternative data sources: International review of current practices" (ESCoE TR-07) by Janine Boshoff, Xuxin Mao and Garry Young.

Outlier detection methodologies for alternative data sources: International review of current practices

Janine Boshoff, Xuxin Mao and Garry Young

Abstract

The construction of consumer price indexes (CPI) has historically relied on manually and centrally collected price data. As point of sale (POS) scanner data and web-scraped data become more accessible, these alternative data represent a rich new source of information to produce consumer price information. While outlier detection methodologies are well established for traditional data sources, more research is required to better understand the unique quality and format of the alternative data. Several national statistical institutions (NSIs) have already started to conduct research into alternative data source and the outlier detection methodologies that are necessary before these data can be incorporated into CPI calculations. This project reviews the outlier detection methodologies adopted by NSIs that have started to incorporate alternative data sources in their calculation of CPI.

Keywords: consumer price index, multilateral indices, outlier detection, scanner data, web-scraped data.

JEL Classifications: C43, E31

Contact details

Corresponding author: Xuxin Mao, x.mao@niesr.ac.uk

Introduction

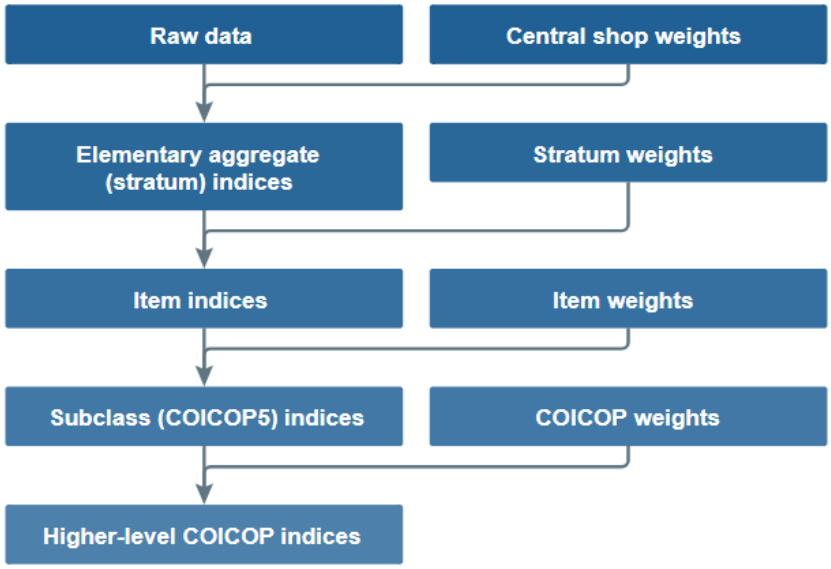
The Office for National Statistics (ONS) hopes to incorporate alternative data sources into the production of aggregate measures of consumer price statistics by Quarter 1 (January to March) 2023 (ONS, 2019a). The increased availability of web-scraped and point of sale scanner data represent a rich data source that could supplement current data collection methods. These alternative data sources have a unique structure and therefore require different methodologies for the detection of outliers that could influence consumer price statistics.
For this reason, ONS have commissioned research on outlier detection methods for web scraped and scanner data to provide an overview of the procedures employed by other national statistics agencies. The project is limited to a review of published literature and ongoing working papers by statistical agencies that the ONS have selected for consideration.

Section 1 will provide a summary of ONS procedure for outlier detection as used in the current construction of consumer price statistics and provide highlight the benefits associated with alternative data sources. Section 2 sets out Eurostat guidelines on outlier detection methodologies, while section 3 provides an overview of outlier detection methods used by other statistical agencies. The literature review concludes with a summary table and provides ONS with an overview of the procedures used in other statistical agencies when incorporating alternative data sources.

1. Overview of current ONS procedures

Figure 1 provides an overview of the aggregation procedures used in calculating the Consumer Price Index including owner occupiers housing costs (CPIH). Except for some centrally collected items and prices for fresh fruit and vegetables, the raw data undergoes two phases of validation checks to ensure consistency in price collection.

Figure *1*: Aggregation procedure in the CPI



Source: ONS (2019b)

Phase one involves a min-max methodology to detect any outliers in the raw data. A minimum and maximum range is derived for each item based on a valid, non-zero price quote recorded in the previous month. The smallest and largest prices are then expanded by price range percentages to form the min-max range. The price range percentages are set by item groups and applied to both the price level and the price change.

Using only the prices recorded as valid for the current month, phase two uses the Tukey algorithm to identify additional outliers. The algorithm calculates the price relative for each individual price quote and sorts these into ascending order. A trimmed mean is calculated after removing the top and bottom 5% of the price relatives, and upper and lower "midmeans" are calculated based on the sub-samples that fall above and below the trimmed mean. The upper and lower Tukey limit is calculated as the trimmed mean plus/minus the difference between the trimmed mean and the midmeans. Price levels or price relatives that fall outside

of the Tukey limits are flagged and validated manually. In the event they do not pass manual validation they are then removed.

Prices that passed the Tukey algorithm and prices that were manually or automatically accepted in phase one are used to calculate preliminary item indices. All prices that fail the Tukey algorithm but have price relatives that fall within 10 index points of the item index are considered valid and used for calculating the elementary aggregates.

While these outlier detection methods are sufficient for locally and centrally collected price quotes, the volume and structure of alternative data sources call for further investigation of potential methodologies to address outliers.

Alternative data sources provide many benefits compared with more traditional methods of data collection, including improved product coverage, high frequency of collection, as well as potential cost savings. There is also potential to provide greater regional coverage of prices and expenditure such as, for example, regional inflation measures. For this reason, this literature review considers the use of alternative data sources in other statistical institutions and the outlier detection methodologies employed by these institutions.

2. Eurostat guidelines on outlier detection

While Eurostat does not have a consolidated view on outlier detection methodologies for alternative data sources, the institution refers to several documents for guidance on best practice across four different groupings.

Data validation and Outlier detection in standard CPI compilation

For accepted methods in outlier detection used in standard CPI compilation, Eurostat references the International Labour Organisation's (ILO) January 2020 draft Consumer Price Index Manual, which suggests the following methods:

i. Statistical agencies can use the **median and quartile values** of the price ratios in the sample to determine acceptable limits for the dataset. Based on the assumption that the observed price changes are normally distributed, the limits are usually defined as some multiple of the range between the median and quartiles with observations outside this range flagged as outliers or potential errors. Given that most distributions are skewed, the CPI Manual also suggests three modifications to remedy this:
    a. Transforming price ratios to reflect the transformed distance (S_i) to make the calculation of the distance from the centre the same for both price decreases and increases;

b. Where quartiles lie close to the median, the CPI manual suggests removing observations with no price movement before setting some minimum distance (e.g. five per cent) for monthly changes; and
c. In the event of small sample sizes, the CPI manual suggests combining several samples of similar elementary aggregates before conducting outlier analysis.

ii. The **Tukey Algorithm** is useful in the event when data validation must be conducted with a sample that has many observations with no price change. After sorting the price relatives, the highest and lowest five per cent are flagged as potential outliers and removed from further calculation. After excluding all observations with no price movements, the arithmetic mean of the remaining observations are calculated before splitting the Tukey sample into upper- and lower-midmeans. These midmeans are then the basis of calculating the limits for potential outliers.

iii. Finally, the use of plot charts can be especially helpful in visualising potential outliers for further data validation.

Thresholds for outlier detection in alternative data sources

Eurostat reference the research produced by Van der Grient and de Haan (2010) for the use of thresholds and dumping filters to detect outliers in datasets. Monthly unit values are subjected to two procedures:

i. First, the data are examined based on a threshold calculation. Month-on-month price changes that exceed a factor of 4 are considered outliers and removed from the dataset; and
ii. Second, the data are subjected to a dumping filter where an algorithm determines whether an observation exhibits a significant price decrease in conjunction with a steep drop-off in expenditure numbers. A dumping filter is most useful in the case of stock clearances where items are sold at exceptionally low prices. Given that the item will no longer be available for purchase in the future, it is removed from the dataset to offset the downward bias it introduces to subsequent index calculations.

Impact calculations to detect outliers

A novel approach to detecting outliers in scanner data is the use of impact calculations as developed by Webster and Tarnow-Mordi (2019). Their research develops a methodology to decompose multilateral price indexes into contributions from individual commodities to understand the aggregate contribution a commodity makes to price index movements.

One feature of multilateral indexes is that the price comparison between two time periods could depend on prices in other periods, and on whether commodities are sold or not in either of these time periods. Without decomposing multilateral index movements, it would be difficult to determine which commodity price change has the greatest impact on the price comparisons. Webster and Tarnow-Mordi (2019) decompose three popular multilateral methods to illustrate the functionality of their impact calculations: the Time Product Dummy (TPD), the GEKS method and the Geary-Khamis (GK) method.

Using a modified data set from Turvey (1979), their illustration contains monthly price and quantity information related to five fruit commodities over four years. Apples, grapes and oranges are sold every month of the year while the seasonal fruits (peaches and strawberries) are only sold for a few months of the year. Constructing the TPD, GEKS and GK indexes, it becomes evident that the month of May is subject to steep price increases. Webster and Tarnow-Mordi (2019) deconstructed the commodity contributions for the price change between April and May 1973, and found the following interesting results:

- Strawberries, a seasonal commodity with an intermittent sales pattern, contributes to the aggregate increase in prices in May 1973; and
- The TPD and GK decompositions indicate that although some commodities have price increases between April and May, these could contribute to an aggregate price decrease since the commodities contribute less than one. Contributions depend on both weight and prices, and a change in weight can sometimes offset increases in price

The impact change method is most useful to organisations constricting multilateral indexes using scanner data, as it allows for the decomposition to account for changes in both price and quantity sold.

3. Review of current procedures used by statistics agencies

3.1. Australian Bureau of Statistics (ABS)

The Australian Bureau of Statistics (ABS) has been using scanner data meaningfully in the compilation of the CPI since the first quarter of 2014 (ABS, 2013). A 'direct replacement' approach was used where unit values calculated from the scanner data was combined with manually collected prices at the elementary aggregate (EA) level and accounted for approximately 25 per cent of the CPI weight (ABS, 2016).

Realising that scanner data could be used to further enhance the CPI, the ABS conducted additional research and published an implementation plan that outlined the methodologies considered to compile 28 expenditure classes (EC) in CPI (ABS, 2017).

In order to implement new methods using scanner data, a statistical agency must make two important distinctions in terms of product definition and elementary aggregation (Dalén, 2017):

i. Research from other national statistics agencies found that **product definition** becomes especially important when using scanner data, as matched model multilateral indexes are prone to downward bias if items are 'relaunched' using different product identifiers (Chessa, 2016). To remedy this, ABS uses the stock keeping unit (SKU) of products instead of the product barcodes as the broader category captures these changes better. This definition of a homogenous product is also consistent with other methods of price collection in the CPI (ABS, 2017).

ii. ABS had initially opted for an **elementary aggregation** that allowed for modified price aggregation directly to the EC level using a multilateral method (ABS, 2016). Further research, however, indicated that the performance of the multilateral method could be improved if it were applied below the EC level. This form of elementary aggregation is consistent with practices used at other statistical agencies (Chessa, 2016; Dalén 2017).

A review of four multilateral methods[1] to compile the Australian CPI led the ABS to conclude that the GEKS-Törnqvist multilateral method is sensitive to products that exhibit clearance prices. Therefore, the ABS define any products with periods characterised by atypical prices and very small sale quantities as outliers, and these observations are removed from the

[1] The 2016 Information paper reviewed four potential multilateral methods: 1) Weighted Time Product Dummy (TPD) method; 2) Geary-Khamis (GK) method; 3) Quality adjusted unit value using TPD (QAUV_TPD) method and 4) GEKS-Törnqvist method.

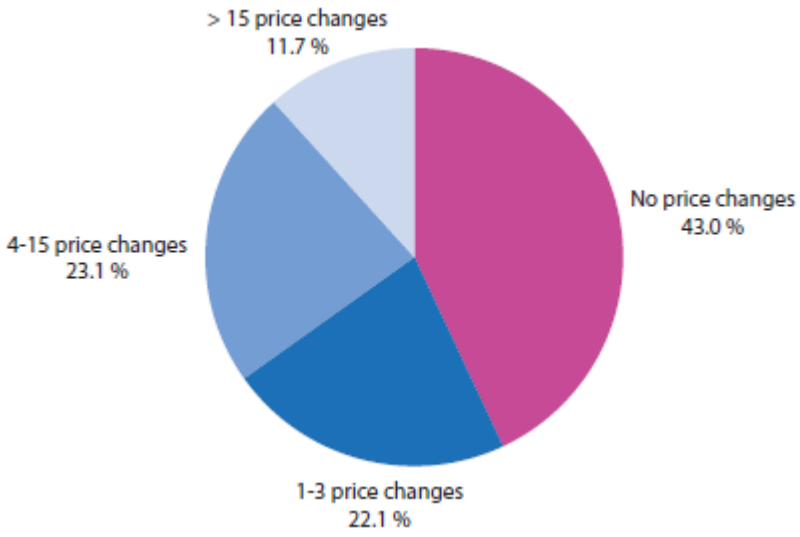
dataset. The exclusion of clearance prices is consistent across all methods of data collection (ABS, 2017).

Thereafter, the choice of time aggregation limits the impact of outliers on price calculation as more aggregation allows for the smoothing of both price and quantity bounces observed in the data. In line with research, ABS apply quarterly aggregation to its scanner data to align with its publication frequency of the CPI (Diewert, Fox and de Haan, 2016; ABS, 2017).

3.2. German Federal Statistical Office (Destatis)

Research by Blaudow and Burg (2018) indicated that alternative data sources have an increasingly imported role to play in the production of the CPI in an environment with dynamic price setting: if prices change more than once per month, traditional methods for price collection could include less representative prices instead of the true unit value price in a month. In their study, which web-scraped price data from 14 online shops between December 2016 to March 2017, some 57 per cent of online prices changed once a month or more.

Figure *2*: Share of products by number of price changes



Source: Blaudow and Burg (2018)

Destatis already uses web-scraped data in its production of the CPI and are now conducting further research to incorporate scanner data into the monthly production. Outlier detection methodologies differ for each of these sources and will be discussed separately.

Web-scraped data

Destatis uses two methodologies to detect outliers on their web-scraped dataset:

i. First, the percentage change between two observations (t and t-1) are calculated and any percentage change greater than factor five is flagged. Thus, items that exhibit current prices 400 per cent higher or 80 per cent lower than the price in the preceding observation will be removed from the dataset; and
ii. Second, confidence intervals are calculated for the complete time series over the full observation time horizon, and any data that falls outside the confidence interval by a factor of five are also flagged as outliers and removed from the dataset.

Scanner data

While scanner data is not currently used in the production of the monthly CPI, Destatis currently receives data from one supermarket chain that accounts for a small proportion (<30%) of the market share. The scanner data cover products in the food and non-alcoholic beverages and alcohol and tobacco consumer segments. The outlier methodology involves three checks:

i. First, the data is reviewed and any observations with turnover or sales quantities less than zero are removed from the dataset;
ii. Thereafter, percentage changes from the previous period (t and t-1) are calculated for turnover information, sales quantities and the number of items offered. Any sales quantities that exhibit increases or decreases of 20 per cent or more are flagged as potential outliers. There is also a comparison of item levels between the two observation periods and evaluated against a graduated threshold. Any observations flagged as potential outliers are then removed from the dataset; and
iii. Finally, Destatis conducts a consistency check across two variables based on unit of measure and net weight. A preliminary unit value is calculated at the item level (turnover divided by sales quantity) to verify the consistency between sales quantity and turnovers. Should the unit values fall below 0.1 or above 2,000 they are flagged as outliers and removed from the data.

3.3. Hungarian Central Statistical Office (HCSO)

The Hungarian Central Statistical Office (HCSO) is currently conducting research into alternative data sources, but as yet have not incorporated these data sources into the production of their Consumer Price Index (CPI). The overview provided below is based on two reports published by the HCSO in December 2019.

The HCSO has an agreement with a product comparison webpage (Arukereso[2]) to collect information on electronic devices using a web scraping software developed in-house. Although HCSO and Arukereso have had a standing agreement about scraping for several years, IT issues resulted in a loss of data. For this reason, their report on alternative data sources was limited to data available from 1 January 2017 to 31 December 2017.

During 2017, HCSO selected four electric devices for which it would scrape data: tablets, microwave ovens, electric kettles and refrigerators. Due to the variety of product types available on the website, HCSO filtered the data to "representative items" listed in the current CPI price collection for comparability purposes. The filtering also allowed for ±10% differences in the characteristics of the representative goods, e.g. tablets with display sizes between 6.3 and 8.8 inches were collected if there were not tablets with 7 or 8 inch displays available. In addition to price information, the software also scrapes supplementary information on product characteristics[3] such as brand, power dimensions, electrical specifications, product attributes and the shop at which the item is available for purchase. Table 1 provides the representative items for which information was collected.

Table *1*: Representative products for which information was web scraped

Refrigerators	Single door, A+ or A++ energy class, approximately 120 litre capacity with built-in freezer
Combined refrigerators	Two doors, A+ or A++ energy class, approximately 200-225 litre capacity with 75-100 litre freezer capacity
Microwave ovens	17-23 litre capacity, 700-1000-watt power, without grill function
Tablets	Android® operating system, approximate 7-8" display, 16-32 GB memory
Electric kettle	1.5-1.7 litre capacity, 2000-2400-watt power

Source: Hungarian Central Statistics Office (2019b)

After the data has been compiled, HCSO starts by examining the distribution of prices to check for any obvious errors. The first step is to manually delete any observations that lie well outside the normal price range for a product, e.g. tablets offered at a price of 10 million Hungarian

[2]Arukereso is the market leading price comparison website in Hungary. Developed in 2004, it now hosts 2,974 retailers and more than 12 million products on its website.

[3]HCSO conducted research into using web-scraped price data in hedonic regression models for the purpose of quality adjustments to HICP estimations.

Florint (HUF) (approximately €31,000) while the average price is 70,000 HUF (approximately €220).

Thereafter, the data are subjected to an automated outlier detection method that produces a series of warnings to determine whether an observation qualifies as an outlier. The process involves a series of calculations with binary outcomes that funnel the data onto the next calculation, assigning flags to certain outcomes. Figure 3 provides a graphical representation of the process described below.

First, the method examines whether the current price (p_t) is the same as the price recorded the day before (p_{t-1}):

- If the price is unchanged, the system applies the same flags to the current price as what was recorded for p_{t-1}.
- If the price is different from that recorded to the previous day, the process moves on to the next step.

Second, the method examines whether the observation is a new product:

- If the observation is a new product, the system produces an adjusted boxplot based on the exponential model developed by Vanderviere and Huber (2004).
- If the observation is not a new product, the process moves on to the next step.

Third, the method determines whether the change in price falls within an acceptable interval [a, b][4]:

- If the price observed falls outside of the acceptable interval, the data gets flag 1 assigned to it.
- If the observation falls within the acceptable interval, the process moves on to the next step.

Fourth, the process determines whether the ratio of the new price and the average price in the previous day falls within an acceptable interval [c, d][5]:

- If the ratio falls outside of the interval range, the data is assigned flag 2.
- If the ratio falls within this acceptable range, then the process moves on to the next step.

[4]Parameters are chosen based on the distribution of the examined phenomena.

[5]Parameters are chosen based on the distribution of the examined phenomena.

Fifth, for all relevant observations the system produces an adjusted boxplot based on the exponential model developed by Vanderviere and Huber (2004).

- If the observation falls outside of the adjusted boxplot range, the data is assigned flag 3.
- If the observation falls within the adjusted boxplot range, then the process moves on to the next step.

Finally, the process determines that any observations that have 2 or more flags is an outlier which is subsequently removed from the final dataset.

The HCSO have used an adjusted boxplot method in order to analyse the distribution of the price data daily. The price data is divided into quartiles (Q_1, Q_2, Q_3) and the interquartile range (IQR) is calculated as IQR = $Q_3 - Q_1$. To measure the skewness of the distribution, the medcouple (MC) is defined as:

$$\text{MC} = \text{med}\left[\frac{(p_j - Q_2) - (Q_2 - p_i)}{p_j - p_i}\right]$$

Vanderviere and Huber (2004) extended the original boxplot by introducing two functions h_l(MC) and h_r(MC) within the cut-off values to classify outliers. Therefore, the standard interval

[Q_1 – 1.5 IQR ; Q_3 + 1.5 IQR]

is augmented to define the boundaries of the interval to be:

[Q_1 – h_l(MC) IQR ; Q_3 + h_r(MC) IQR]

The use of two different functions allow for the whiskers to be different lengths, and in the even that $h_l(0) = h_r(0) = 1.5$, then the boxplot reverts to the original interval range for symmetric distributions. Under the exponential model, the two functions become:

h_l(MC) = 1.5 e^{gMC} ; h_r (MC) = 1.5e^{fMC}

In order to determine the constants used in the exponential model, Vanderviere and Huber (2004) require the expected percentage of marked outliers to be equal to 0.7 per cent, in line with the outlier rule for a normal distribution boxplot. For the exponential model, the two augmenting functions are then calculated as follows:

$$\text{ln } ⅔ \left[\frac{(Q_1 - Q_{lowermid-mean})}{IQR}\right] = f\,\text{MC}$$

$$\text{ln } ⅔ \left[\frac{(Q_{uppermid-mean} - Q_3)}{IQR}\right] = \text{g MC}$$

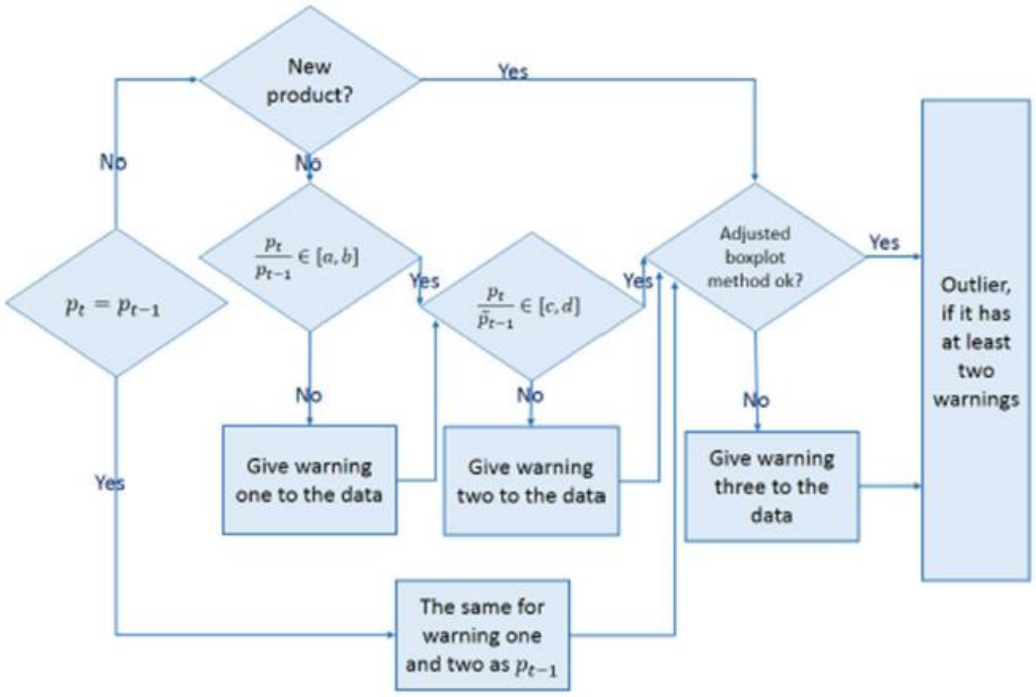
Therefore, if the medcouple is greater than or equal to zero (MC ≥ 0) then the acceptance interval becomes:

$[Q_1 - 1.5\ e^{-fMC}\ IQR\ ;\ Q_3 + 1.5\ e^{gMC}\ IQR]$

If the medcouple is less than zero (MC < 0), then the acceptance interval becomes:

$[Q_1 - 1.5\ e^{-gMC}\ IQR\ ;\ Q_3 + 1.5\ e^{fMC}\ IQR]$

Figure *3*: Outlier detection process used for web-scraped data



Source: Hungarian Central Statistics Office (2019b)

Based on the initial research, HCSO found that approximately 10% of observations are flagged, with only 5% of all the observations finally determined as outliers.

3.4. Statistics Belgium (Statbel)

Statistics Belgium (Statbel) is one of the few institutions that have already incorporated scanner data and web-scraped data into the production of their CPI measure and are now conducting further research into extending its use of web-scraped data.

Table *2*: Consumer segments with alternative data sources

Scanner data	Food and non-alcoholic beverages Alcoholic beverages and tobacco Miscellaneous small tool accessories	Non-durable household goods Products for pets Paper products Other stationery and drawing materials
Web-scraped data	Clothing Footwear Hotel reservations Airfares International train travel Second-hand cars Consumer electronics	Drugstores Books Videogames DVD & Blu-ray discs Supermarkets Student rooms

Source: Adapted from Van Loon and Roels (2018)

Scanner data

Statbel has incorporated scanner data from the three largest supermarket chains into the construction of the CPI since 2015. These chains account for approximately 80 per cent of the market and provide information on product groups related to food and non-alcoholic beverages (FNAB), alcoholic beverages and tobacco, personal care products and other household goods (Van Loon and Roels, 2018).

Indices calculated using scanner data use a dynamic basket methodology with a monthly chained Jevons index. This dynamic basket uses turnover figures in two adjacent months to determine if the product is included in the sample, subject to the following threshold calculation:

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

Where s_m is the market share of each matching product in month m, s_{m-1} is the market share of each matching product in month m – 1, n is the number of products and λ is 1.25. An outlier filter excludes any products that show extreme price changes between two months, but Van Loon and Roels (2018) note that this usually applies to items that are obtained for free with loyalty cards. Two additional dumping filters are also applied to the data to control for stock clearances:

- The first filter excludes all observations that show a sharp decrease in both price and sales quantities; and
- The second filter excludes any products that exhibit a sharp decrease in quantities sold while the price remains relatively stable.

Homogenous products are defined using the supermarket chain SKUs and GTINs related to relaunches and replacements are linked to related products using data and text mining. These data are then used to calculate an index for each retailer at the ECOICOP 5-digit level. Indices at the level are then combined with other data sources using a stratification model using the retailer's relative weight in the market share after which an elementary level index is calculated using the Jevons formula (Van Loon and Roels, 2018).

Following research by de Haan, Hendriks and Scholz (2016) Statbel realised that its current methodology for calculating the CPI with scanner data using monthly chained indexes could introduce chain drift. For this reason, the institution is currently researching multilateral methods[6] that they hope to incorporate in their CPI production during 2020.

[6] The multilateral methods currently under review include the Geary-Khamis and augmented Lehr method, the Time Product Dummy method and the GEKS-Törnqvist method. Statbel is also

Web-scraped data

Web-scraped data for international train travel and videogames is already incorporated into the calculation of the CPI, but Statbel is currently running 70 web-scraping scripts daily to collect data on other consumer segments for further research (see table 2).

Initial research suggests that each consumer segment is currently subject to different outlier detection methodologies, and Statbel has yet to develop a generic methodology to apply across all web-scraped data. For example, in the footwear segment potential outliers like promotional prices are included in the sample during traditional sales periods but excluded otherwise (e.g. "flash sales"). The dataset for second-hand cars are subject to outlier detection methodologies that exclude salvaged or damaged cars, or cars that are earmarked for export; thereafter monthly aggregation limits the impact of outliers in the dataset (Van Loon and Roels, 2018).

3.5. Statistics Netherlands (CBS)

Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS) has been one of the leading authorities to experiment with the use of scanner data in the production of the monthly CPI. Scanner data was initially introduced into the Dutch CPI in 2002 using data provided by two supermarket chains (Chessa, 2016). By January 2010, the CBS had access to scanner data from seven supermarket chains and were using data accounting for around half of the market share and contributing 5 per cent of the CPI weight (Van der Grient and de Haan, 2010). Since January 2013, CBS no longer used surveys to collect price data as it had access to data from ten supermarket chains. The dataset has been further extended to include scanner data from other retailers (e.g. DIY stores) and electronic data related to fuel prices, mobile phones and travel agencies, accounting for more than 20 per cent of the Dutch CPI (Chessa, 2016).

CBS developed different methods for the treatment of data from different types of retailers within the Dutch CPI but eventually found that the system became increasingly complex over time. For that reason, Chessa (2016) outlines the development of a generic method that would reduce any methodological differences in the construction in the Dutch CPI and covers three aspects: data processing, product homogeneity and price index calculation.

researching various extension methods for the comparison window: movement splice, window splice, half splice, mean splice, fixed base monthly expanding window and fixed base moving window.

Data processing

The current methodology aims to process all global trade item numbers (GTINs) which includes all assortment changes directly into the index calculation. The CBS uses price and turnover filters to identify outliers and errors within the dataset (Chessa, 2016). Van der Grient and de Haan (2010) provided two examples of these filters:

- An automated procedure identifies month-to-month price changes that exceed a factor of four as implausible and declares these observations invalid. Therefore, if an observation is 300 per cent higher or 75 per cent lower than that recorded in the preceding month, the observation is deleted; and
- An algorithm known as a dumping filter will exclude any items that exhibit significant price decreases in conjunction with a steep decline in expenditures.

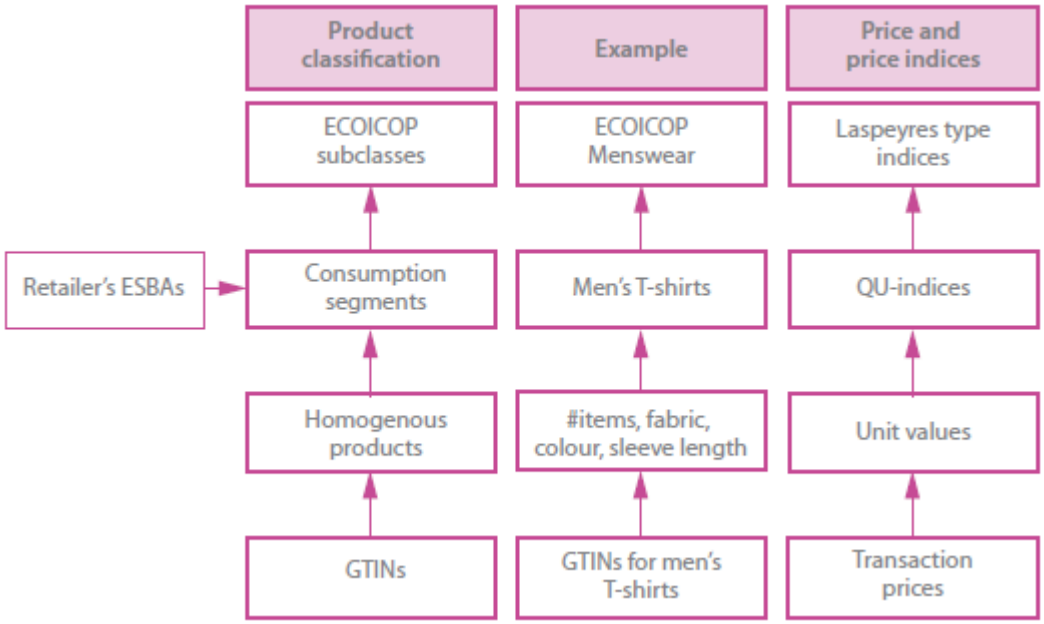
Product homogeneity

Although GTINs represent the highest level of homogeneity, they are hampered by the 'relaunch' phenomenon identified by Chessa (2013). GTINs can, instead, be linked and combined into homogenous products by either:

- Linking old and new GTINs using the retailer's stock keeping units (SKU); or
- If SKUs are not available, GTINs can be combined by matching up item characteristics.

Figure 4 illustrates the use of GTINs and product characteristics used in the creation of a homogenous product category at the level of consumption segment. CBS relies on retailer classification of GTINs ('ESBAs' in the Dutch CPI) to determine a consumption segment below the level of the ECOICOP to link GTINs to a particular ECOICOP (Chessa, 2016).

Figure *4*: Product classification using GTINs and product characteristics



Source: Chessa (2016)

Price index calculation

Once a homogenous product is defined, a unit value is calculated by adding turnover and quantities sold for GTINs that below to this product. These numbers are combined to form quality-adjusted unit value indices (QU-indices) for consumption segments that will eventually be linked to an ECOICOP (Chessa, 2016).

3.6. Statistics New Zealand (SNZ)

Statistics New Zealand (SNZ) incorporated scanner data for consumer electronics into the production of the CPI starting in the September quarter (Q3) 2014. The market research company GfK provides monthly scanner data to SNZ and table 2 provides an overview of the consumer electronics covered in the dataset, as well as their weight in unique consumer segments and the headline CPI, respectively.

Table 3: Scanner data used in the New Zealand CPI

Electronic item	**Category contribution weight**	**Headline CPI weight**
Heat pumps	Major household appliances 20 per cent	0.71 per cent
Cellphone handsets	Telecommunication equipment 90 per cent	0.29 per cent
Desktop computers	Audio-visual and computing equipment 80 per cent	1.16 per cent
Laptop computers		
Tablet computers		
Multi-function devices		
Digital cameras		
Digital camera memory cards		
Television sets		
Set-top boxes for television sets		
DVD, Blu-ray players and player/recorders		
Home theatre and stereo systems		

Source: Statistics New Zealand (2014)

SNZ pre-aggregates the monthly to a quarterly level before deriving the quarterly index that is published. Not only does this minimise the impact of any potential outliers, it also ensures that products sold in each month of the quarter are weighted for price deflation (Statistics New Zealand, 2014).

At the time of publication, SNZ usually only have the first two months data for consumer electronics. The agency uses these two months of data to derive the Imputation Törnqvist rolling year GEKS (ITRYGEKS) index, with the third month of the quarter incorporated in the five-quarter estimation window used to calculate the next quarter's CPI. This allows SNZ to incorporate the timeliest information into the current estimation and allows the use of all available information over the five-quarter estimation window (Statistics New Zealand, 2014).

Agency	Experience using alternative data sources		Goods covered in alternative data sources		Currently used in production of official CPI	Outlier detection method used
	Scanner data	Web-scraped data	Scanner data	Web-scraped data		
Australian Bureau for Statistics (ABS)	▪	▪	Food and non-alcoholic beverages, tobacco, personal care products, household cleaning items, pets and related products, and other non-durable household products.	Clothing and footwear	▪	Exclusion of clearance prices and relevant aggregation to reduce impact of outliers
German Federal Statistical Office (Destatis)	▪	▪	Food and non-alcoholic beverages, alcoholic beverages and tobacco	-	▪	Threshold calculation based on percentage change and confidence intervals
Hungarian Central Statistics Office (HCSO)	✗	▪	-	Electronic devices: tablets, microwave ovens, electric kettles and refrigerators	✗	Adjusted boxplot based on exponential model
Statistics Belgium (Statbel)	▪	▪	Food and non-alcoholic beverages, alcoholic beverages and tobacco, non-durable household goods, paper products, pet product, stationery	Clothing and footwear, hotels, airfares, train tickets, second-hand cars, books, consumer electronics, video games, DVD & Blu-ray	▪	Threshold calculation, outlier filter and two dumping filters

Agency	Experience using alternative data sources		Goods covered in alternative data sources		Currently used in production of official CPI	Outlier detection method used
	Scanner data	Web-scraped data	Scanner data	Web-scraped data		
			and drawing materials, miscellaneous small tool accessories and personal care articles.			

Agency	Experience using alternative data sources		Goods covered in alternative data sources		Currently used in production of official CPI	Outlier detection method used
	Scanner data	Web-scraped data	Scanner data	Web-scraped data		
Statistics Netherlands (CBS)	▪	-	Food and non-alcoholic beverages, Wine, Beer, Tools and equipment (house and garden), Household maintenance goods and services, Medical and pharmaceutical products, personal care goods and appliances, pet care and pet food.	-	▪	Price and turnover filters identify potential outliers within the dataset
Statistics New Zealand	▪	✘	Consumer electronics: heat pumps, desktop computers, laptop computers, tablet computers, multi-function devices, cell phone handsets, digital cameras, digital camera memory cards, TV sets, set-top boxes for TV sets, home	-	▪	Relevant aggregation to reduce impact of outliers

			theatre systems, stereo systems, DVD players, Blu-ray players and player/recorders			

References

Australian Bureau of Statistics (ABS) (2013), *Feature Article: The Use of Transactions Data to compile the Australian Consumer Price Index*. Cat. No. 6401.0. ABS, Canberra.

Australian Bureau of Statistics (ABS) (2016), *Information Paper: Making Greater Use of Transactions Data to Compile the Consumer Price Index*. Cat. No. 6401.0.60.003. ABS, Canberra.

Australian Bureau of Statistics (ABS) (2017), *Information Paper: An Implementation Plan to maximise the use of Transactions Data in the CPI*. Cat. No. 6401.0.60.004. ABS, Canberra.

Blaudow, C. and Burg, F. (2018), *Dynamic pricing as a challenge for consumer price statistics*. Eurostat review of National Accounts and Macroeconomic Indicators, 1, pg. 79-94.

Chessa, A.G. (2013), *Comparing scanner data and survey data for measuring price change of drugstore articles*, Paper presented at the Workshop on Scanner Data for HICP, 26-27 September 2013, Lisbon, Portugal.

Chessa, A.G. (2016), *A new methodology for processing scanner data in the Dutch CPI*. Eurostat review of National Accounts and Macroeconomic Indicators, 1, pg. 49-70.

Dalén, J. (2017), *Unit Values in Scanner Data – Some Operational Issues*. Paper presented at the fifteenth Ottawa Group meeting, 10-12 May 2017, Elville am Rhein, Germany.

De Haan J., Hendriks, R. and Scholz, M. (2016), *A comparison of Weighted Time Product Dummy and Time Dummy Hedonic Indexes*. Graz Economics Papers 2016 (13), University of Graz, Department of Economics.

Diewert, E.W., Fox, K.J. and de Haan, J. (2016), *A newly identified source of potential CPI bias: Weekly versus monthly unit value price indexes*. Economics Letters, 141, pg. 169-172.

Grient, H.A., van der and de Haan, J. (2010), *The use of supermarket scanner data in the Dutch CPI*, Paper presented at the Joint ECE/ILO Workshop on Scanner Data, 10 May 2010, Geneva, Switzerland.

Hungarian Central Statistical Office (HCSO) (2019a), *Weighting of web-scaped price data for improving the quality of HICP in HCSO*.

Hungarian Central Statistical Office (HCSO) (2019b), *Quality adjustment of web-scraped price data for improving the quality of HICP in HCSO*.

International Labour Organisation (ILO) (2020), *Consumer Price Index Manual: Concepts and Methods* (Draft version: January 2020). Available from www.imf.org

Office for National Statistics (ONS) (2019a), *Using alternative data sources in consumer prices indices,* May 2019.

Office for National Statistics (2019b), *Consumer Prices Indices Technical Manual*, 2019.

Statistics New Zealand (2014), *Measuring price change for consumer electronics using scanner data*. Available from www.stats.govt.nz

Vanderviere, E. and Huber, M. (2004), *An adjusted boxplot for skewed distributions*. Paper presented at the sixteenth Computational Statistics (COMPSTAT) symposium, January 2004, Prague, Czech Republic.

Van Loon, K. and Roels, D. (2018), *Integrating big data in the Belgian CPI*, Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018, Geneva, Switzerland.

Webster, M. and Tarnow-Mordi, R.C. (2019). *Decomposing Multilateral Price Indexes into the Contributions of Individual Commodities*. Journal of Official Statistics, Volume 35, No. 2, pg. 461-486.