EUROPEAN COMMISSION
EUROSTAT

Directorate C : National Accounts; prices and key indicators
Unit C-4: Price statistics; Purchasing power parities; Housing statistics

Harmonised Index of Consumer Prices

Practical Guide for Processing Supermarket Scanner Data

Eurostat

September 2017

Table of Contents

For inquires please contact D.J. Hoogerdijk at ESTAT-prc-stats-methods@ec.europa.eu

Foreword

Scanner data is a relatively new data source for consumer price statistics. A fifth of EU countries use such data already, several are likely to start using it within a few years, and others are considering doing so.

When EU countries start using scanner data, they generally start with data from supermarkets. Therefore this practical guide deals with the use of scanner data for the range of food, beverages and personal and home care products generally available in supermarkets, in short: scanner data for supermarkets. This means it will also be useful for chemists' shops and other retailers selling this type of goods.

The guide does not deal with processing scanner data for goods like clothing and electronics, where the assortment changes very frequently.

The methods described in this guide should not be applied to volatile categories of goods like clothing. The methods outlined here assume a degree of stability in the assortments available in supermarkets, which means that the goods' Global Trade Item Numbers (GTINs) or other item codes remain valid for longer periods of time.

Why have a practical guide?

Firstly, it is **Eurostat's responsibility** to ensure that the Harmonised Indexes of Consumer Prices (HICPs) of EU countries are comparable. Using scanner data risks introducing incomparability if national statistical institutes (NSIs) develop their own methods for processing the data. As more NSIs want to start using scanner data, it makes sense to learn from others and align practices with current best practice. This will reduce the risk of incomparability.

Secondly, Eurostat supports the **modernisation of price statistics**, the aim being to ensure that price collection methods remain appropriate in a world of increasingly dynamic markets for consumer goods, dynamic pricing and ingenious ways of providing discounts. The use of scanner data is a partial response to these challenges. This guide is one form of support, in addition to holding regular meetings and workshops.

Thirdly, Eurostat hopes that the guide will give **users** some insight into the complexity of using scanner data for the HICP and the issues involved.

Why now?

A few NSIs have been using supermarket scanner data for many years, and methods have emerged that provide reliable indices for supermarkets. The methods will be described in this guideline[1]. As more NSIs want to start using scanner data, the time seems ripe for a practical guide.

[1] New methods including multilateral ones are currently being explored to process scanner data. These methods and the discussion on them have not yet matured sufficiently for the results to be included in this document.

The second reason is that current legislation does not explicitly deal with the use of scanner data. After the recent adoption of a new framework regulation[2], implementing acts will have to be drawn up that should also cover scanner data. This guide is designed to provide practical background information for the drafting process.

This guide describes the situation in 2017. It will need to be updated as the use of scanner data develops and broadens.

We hope the guide will help speed up the process of using scanner data and ensure comparability among the national HICPs.

[2] Regulation (EU) 2016/792 of the European Parliament and of the Council.

1. Introduction

Scanner data is generated by point-of-sales terminals in shops and provides information at the level of the barcode or, more correctly, Global Trade Item Number (GTIN, formerly known as EAN code). Sales terminals record each transaction. Scanner data, as currently used by national statistics institutes (NSI), is an aggregation of the turnover and quantity of individual transactions per GTIN for a given period and location (outlet or retailer) and provides information on what the product is. This allows a unit value price to be calculated for each GTIN. Codes other than GTINs, dealt with in Chapter 4, can also be used, but this guide will use the term 'item code' throughout.

The scanner data that is available for NSIs is not generated or collected for the specific purpose of compiling consumer price statistics; it is often similar to data used by the retailer and market researchers to monitor market developments. For NSIs, scanner data is therefore a *secondary data source* that is 'data originally collected for a different purpose and reused for another research question'[3]. By contrast, NSIs determine and are responsible for all steps of the price collection process for primary data sources such as traditional price collection.

We define scanner data as 'transaction data obtained from retail chains containing data on turnover, quantities per item code based on transactions for a given period and from which unit value prices can be derived at item code level'. Data sets with item codes and offer prices or web-scraped data relating to offer prices are not considered as transaction data, even though the processing of this data may be very similar to processing scanner data.

Scanner data can be obtained from a wide variety of retailers: supermarkets, pharmacies, do-it-yourself stores, home electronics or clothing shops, and many others. Currently, scanner data predominantly replaces price collection in supermarkets, in particular for food, beverages and personal and home care products found there. As stated above, this guide will restrict itself to these goods.

Scanner data has several advantages over traditional price collection:

- It provides information on the actual expenditure for all item codes sold (by the retailer whose data is used),
- It provides price information on actual transactions over longer periods of time rather than on just one day per month.
- It excludes items not actually sold and includes certain types of discounts.
- It is a better source of information for the inclusion of new items in the HICP than reliance on price collectors.
- It can reduce the administrative burden on retailers and save costs on price collection

Using scanner data holds the promise of improving the quality of the HICP.

However, scanner data also has its drawbacks. One is a greater dependency on the retailers to provide the data. Another is that there are methodological issues that need to be addressed, in

[3] See Hox and Boeije (2005), Data collection, primary vs secondary, Encyclopaedia of Social Measurement Vol. 1, 593-599.

particular with regard to replacements. The advantages and disadvantages will be discussed in more detail in Chapter 2.4.

The guide explains what data to ask for, from whom and with what frequency, and provides guidance on checking the quality of the data and on how to process it. However, it does not address **how** to get the data, as this depends on institutional and legal arrangements at national level, national customs and the negotiating skills of those concerned. The experiences of the NSIs that have implemented scanner data show that maintaining good relations with retailers is essential, to ensure that data continues to be supplied and issues that inevitably crop up are resolved in a timely fashion.

Using and obtaining scanner data from other sources or for other statistical purposes, such as the Purchasing Power Parities, is not dealt with here.

Outline of contents

Chapter 2 compares scanner data with traditional price collection and relates the use of such data to the principles of the HICP, in particular the fixed basket and the need to deal with relaunches and replacements.

Chapter 3 discusses the main sampling dimensions: regions, outlets and time. These define the specific features of the data to ask for from retailers.

Chapter 4 deals with item codes, GTINs and other codes commonly used.

Chapter 5 then deals with obtaining scanner data.

Chapter 6 gives an overview of processing scanner data and Chapters 7, 8 and 9 deal with the processing in more detail: classifying item codes to the European classification of individual consumption according to purpose (ECOICOP), the static and dynamic approach to sampling, calculation of indices and integrating scanner data in the overall HICP.

2. Scanner data and the HICP

This chapter compares scanner data with price data collected in a traditional manner and assesses scanner data in the light of the fixed basket principle of the HICP. The chapter will finish with a discussion of the advantages and disadvantages of using scanner data from supermarkets.

2.1. Traditional price collection and scanner data

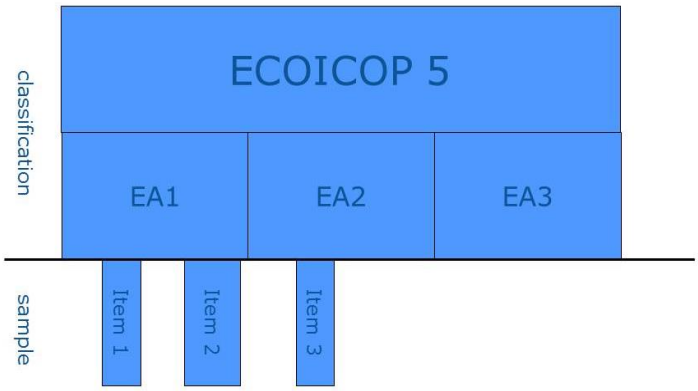
Traditionally, the coverage of the HICP is structured top-down. Total consumer expenditure is broken down using National Accounts data and other sources, like household budget surveys, to lower level ECOICOP[4] aggregates, with each aggregate having its own weight. At the

[4] The European version of the Classification Of Individual Consumption by Purpose. The European version is compatible the official UN version, but has added a more detailed level by breaking up the UN 4-digit classes into 5-digit level subclasses.

lowest level, there are the elementary aggregates below which little information is available on quantities and weights.

Within elementary aggregates product descriptions for items are made for which prices are collected in shops. Product descriptions are relatively broad (e.g. jam, strawberry, 150 – 300 grams) to ensure that one and the same item description can be used for a reasonably long period of time and across different retailers. Figure 1 summarises the structure.

Figure 1: Traditional Structure



EA = elementary aggregate

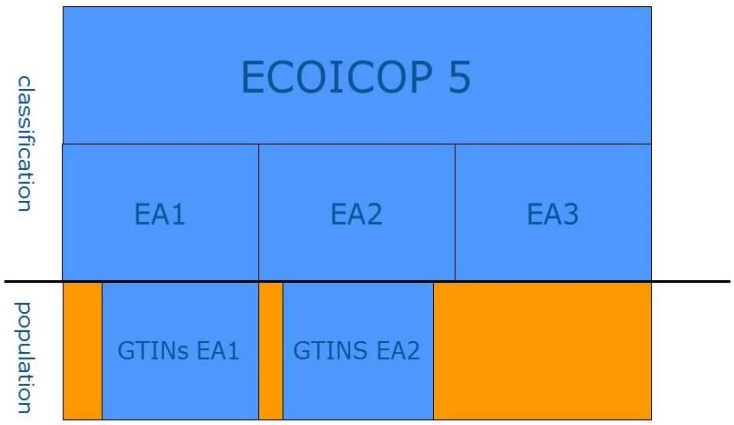
The purposive (selective) sampling of items and the art of making item descriptions is based on common sense and experience. Traditional price collection involves collecting prices of at most a few hundred products[5] in supermarkets each month for food, beverages and other daily necessities, such as personal and home care products. With the exception of products with volatile prices, prices would be collected once a month and only for a sample of outlets. The prices that are collected are **shelf prices**: the prices displayed on labels on the shelves where the items are offered.

Scanner data, for a specific retailer and time period, is an exhaustive listing of all item codes sold, their turnover and the quantities sold.

Scanner data provides the index compiler with all the transactions of a retailer or outlet. Typically, 10 000 – 25 000 item codes are used in a supermarket to cover food, beverages and other daily necessities. Scanner data allows the price statistician to use what was *actually* sold and to include many more items in the HICP than is feasible with traditional price collection. It also means that individual items can be weighted as turnover information is available. Figure 2 summarises the structure for scanner data. Note that, in comparison to the structure in the traditional situation, the only difference is the bottom level. In the case of scanner data not all item codes belonging to a given elementary aggregate need to be used (see Chapter 8), but it is clear exactly what is being left out (the orange part).

[5] Products, as the term is used here, often refers to multiple items. Toothpaste is a product that contains many items (brand and type of toothpaste). See also Figure 3.

Figure 2: Scanner data



Item codes identify a good very precisely, so that two goods with the same item code are identical[6] from the consumer's point of view. The resulting unit value price per item code is the average of prices ***actually*** paid by consumers, including any taxes less subsidies on the items, and after the deduction of discounts from standard prices or charges. Scanner data does not contain prices for ***product offers***[7] i.e. the shelf price at which the product is offered to the consumer.

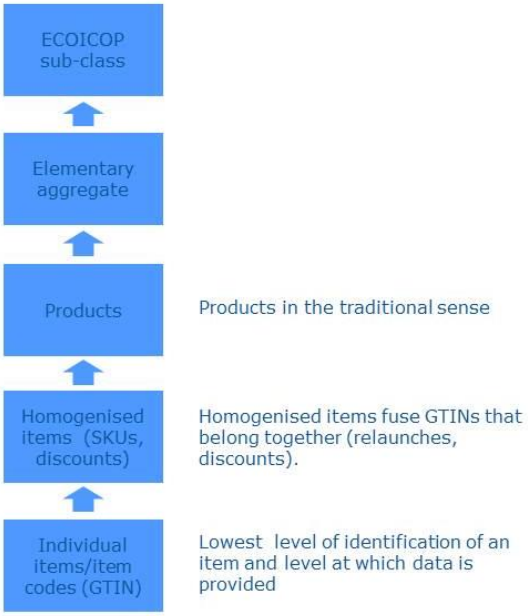
Scanner data reflects the dynamics of actual purchases in each elementary aggregate because each transaction is recorded. The entry of new item codes, the disappearance of item codes and the shifts in the relative importance of items are visible in the data set. We will call this '***churn***'. The disappearance of items is known as ***attrition***, and the rate varies across countries from 25% to 60% of item codes per year. In addition to the genuinely new items introduced, items are often replaced by new versions called 'relaunches'. These are essentially the same item, but with some superficial difference, such as packaging or a new item code. Likewise, discounts (e.g. 20% more content for the same price) are also assigned a new item code. In other cases, replacements are more substantial, for example when items of a certain brand are replaced by similar items of another brand.

Where traditional product descriptions are somewhat broad, item codes in scanner data can be too narrow. To properly include relaunches and discounts, several item codes may need to be taken together. Such codes together form homogenised items, for example strawberry jam of brand X, 250 grams. Similar items together form products, for example all jams of brand X. Such products could then be assigned to the elementary aggregate 'jam' that is part of the ECOICOP 01.1.8.2 (Jams, marmalades and honey). Figure 3 gives the structure.

[6] This point is not always valid for stock-keeping units (SKUs), see section 4, but this does not invalidate the point made here.

[7] The term 'product offer' is specific to the HICP.

Figure 3: Aggregating GTINs to ECOICOP



This guide concerns supermarket scanner data, where the churn is relatively limited compared to clothing. The methods developed in this guide assign item codes directly to elementary aggregates, and not first to a homogenised item. Special care must therefore be taken to capture relaunches and discounts if and when they do occur.

2.2. Scanner data and the fixed basket

The HICP is a Laspeyres-type index. The weights of elementary aggregates, the fixed element, remain unchanged throughout the year. The continuously changing population of item codes mostly affect the level below the elementary aggregate. Hence, the use of scanner data does not infringe on the Laspeyres principle underlying the HICP; actually scanner data could allow the level of elementary aggregate to be taken down to lower levels as information on weights is available. What has changed with the use of scanner data is that the choice of the items within an elementary aggregate is not left to the price collector, but can be based on objective criteria such as actual quantities sold and/or turnover.

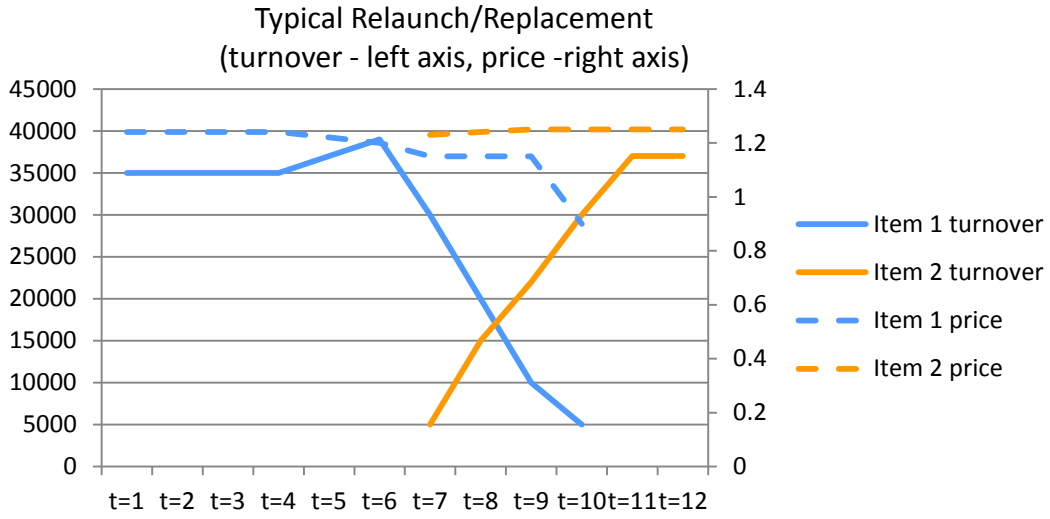
In the traditional system, without any information below the level of elementary aggregate we assume that there is an item described as: Jam, strawberry, 150 – 300 grams. Given that supermarkets always sell jam and strawberry jam is always represented it is a reasonable choice. The question is: of the different varieties sold which one to take? The answer is the most representative one, but how does this help the price collector? Does he or she ask a shop attendant, every time they go to the shop for price collection? And how would the shop attendant know? The answer is that someone should have checked what the most sold strawberry jam is. Scanner data provides the needed information.

In scanner data the representativity of the sample is maintained for an elementary aggregate by letting the processing system, possibly automatically, select the most sold item codes (that refer to strawberry jam). In traditional price collection, price collectors have to trust intuition and common sense and it may happen that prices are collected as long as the item is available even though it is no longer representative. In scanner data the representativity is guaranteed.

The opportunity scanner data offers to solve the issue of replacements lies in the fact that it contains more information on what is happening in the market. Scanner data makes replacements visible, where in traditional price collection these may have remained hidden: price collectors only see what is available in supermarkets, not what is representative. For scanner data it is a challenge to design systems for detecting replacements at item code level, especially in automated production systems.

Replacements and relaunches often follow a certain pattern that is illustrated in the graph below. As one item is replaced by a new item, with a slight price increase, the turnover and quantities sold of the old item decrease, while those of the new item increase.

Graph 1: Replacement



Graph 1 demonstrates the advantage scanner data offers by providing historical information, and at the same time it creates a challenge: for which period should a replacement be made? The price statistician does not have the luxury of hindsight.

In such cases, the turnover and quantity data of two or more item codes could be aggregated into a temporary, fictitious, code and a unit value calculated over both item codes. Replacements should be quality corrected if needed, for example if there is a small difference in the quantity of the contents, e.g. 186 grams compared to 180 grams.

To make sure that the sample remains representative, items should not only be present in the new month but also be representative, i.e. have sufficient turnover or be sold in sufficient quantities. One NSI monitors item codes that sell less than half the number sold in the price reference month, December, and whenever necessary or possible, replaces the item code if a substitute can be found. In this case, item codes are replaced when they fall below a certain threshold. Another NSI ensures the sample is representative by including items that have a certain average turnover over a number of preceding months, together with a specific ranking in the elementary aggregate. The first system 'pushes out' less representative items, whereas the second system 'pulls in' more representative ones.

Seasonal items are treated differently in these systems, taking into account cyclical patterns and non-availability in certain months (see Chapter 9.2).

For supermarkets, package size adjustments may be needed if the brand and type of product remain the same and direct comparisons are to be made. In other cases, a unit value that temporarily combines two item codes or bridged overlap may be used if this gives a correct overall price development.

2.3. Scanner data for weights and other uses

Scanner data can also be used to supplement existing sources of information for weights. The approaches described in this guide do not use weights below the level of elementary aggregate which keeps the methods consistent with traditional methods where prices are not weighted.

2.4. Scanner data: advantages and disadvantages

The advantages come with challenges for the NSI's organisation, the statistical process and the outcomes. A **first challenge** is that the dependence on a retailer is greater than in the past, when a retailer only had to give a price collector permission to visit the outlet. The actual price collection was the responsibility of the NSI and it had full control over it. With scanner data, the dependence on the retailer increases substantially. A common reaction is, firstly, to improve the relationship with the retailers and, secondly, to enter into legal contracts with the retailers that stipulate the details of the contracts.

A **second challenge** is the amount of data that has to be processed and for which the appropriate IT structure and staff needs to be available. For staff this means that more IT-oriented skills are required and, depending on how the traditional price collection is done, there may be a need to have more staff in-house.

The **third challenge** comes from objections to using scanner data because current HICP regulations and recommendations are seen as not accommodating the use of such data. Such objections should be evaluated carefully, as the regulations were not drawn up with scanner data in mind. The processing of scanner data suggests that the period for which missing item codes are imputed should be extended to 14 months to allow for seasonal items to automatically be included when they come in season, but also to allow for shifts in the season between years (see also Chapter 9.2). This extension is not compliant with existing legislation. However, traditional item descriptions are broader and do not equate to item codes. New item codes (possible relaunches and replacements) are included automatically if the dynamic method is used. Item codes without sales are left in the system to ensure seasonal items are automatically included when they come back in season. The implementing acts that will be drafted in the near future will have to take the use of scanner data into account.

The **last objections** could come from ingrained habits and beliefs. Scanner data replaces a long tradition of sampling, price collection, and ways of validating the collected prices. Using scanner data requires a different, new mind-set and it may not always be easy to change people’s habits. Given the limited number of prices collected using traditional methods, it was and is possible to check every single price. This can lead the people who collect the prices to be extremely meticulous and concerned with detail. Scanner data cannot be checked at the same level of detail.

A cost/benefit analysis should be carried out, and the outcome might be different for each NSI.

3. Sampling dimensions

The starting point when deliberating whether or not to use scanner data depends on the answers to the question: is the appropriate data available at the required level of granularity?

This chapter focuses on two dimensions of the sampling process: the where and when. The third dimension is the what: the item code. This will be the subject of Chapter 4.

The HICP is a monthly index at some level of aggregation with regions and outlets or types of outlets[8] appropriately weighted.

3.1. Regions

Larger countries may calculate and publish regional price indices which require collecting scanner data at a regional level.

If the traditional sample is restricted to specific regions or locations, scanner data could be limited to those same regions or locations. However, it may also be possible to combine a sample of regions or locations for non-scanner data items with scanner data aggregated to a national level. This seems a viable option if retailers use national pricing. If retailers use national pricing, the retailer could possibly aggregate scanner data over all regions, but the NSI will regularly need to check whether national pricing is still valid.

If a retailer offers *different* items across regions or uses regional pricing, scanner data should be aggregated by region. The reason is that if data were aggregated to a national level, items sold in *all* regions would have a higher chance of being included in the calculation than items that are important in only one region. This is especially important if regional price indices are made. To give an example: imagine a country where regional producers of dairy products have a substantial market share in particular regional markets. If data is aggregated to national level and used for all regions, regionally important products may be excluded from the regional price development because they are not important at national level.

3.2. Outlets

Many retailers operate several types of outlets, e.g. Carrefour operates Carrefour Hypermarket, Carrefour Market, Carrefour Express and its online stores. The level of service provided, in terms of opening hours or the range of goods available, for instance, often differs from one type of outlet to another, and this may be expressed in the price levels. There is no guarantee that changes in price levels between different outlet types of a retail chain are synchronised.

[8] It is possible for outlet types (supermarkets, bakers, butchers) to be weighted and/or for individual outlets to be weighted.

If a sample of outlets is taken from such a retailer then turnover and quantity data should not be summed over the different types of outlets of the retailer. The data should be summed only over each type of shop and then aggregated to the retailer level. This will allow for a correct replacement of outlets in the sample[9].

Scanner data could be summed over all outlets of a retailer as long as the types and number of outlets per type remains relatively stable. However, large changes in the population of outlets (closures, openings, change of formula or retailer) could affect the index in which case it could be appropriate to process the different types of outlets separately.

If outlets are free to set their own prices, then scanner data should be supplied at the outlet level.

3.3. Time

In the traditional way of collecting prices on one day in the month this is considered a sufficient approximation of the monthly average price. Products with volatile prices, such as vegetables, are an exception, and prices are collected more often. Scanner data offers the opportunity to process data relating to longer time periods than a day. Currently scanner data is commonly aggregated over a week and delivered and processed per week.

The time period over which scanner data is aggregated should align with the retailer's pricing policy (discounts), thereby enabling price changes to be monitored.

If the day of the week or time of the day is a quality aspect of the product and the price is set accordingly, then this aspect could be taken into consideration. However, if this is a cyclical pattern ('prices are always higher at the weekend'), then if the same 'cycles' are included each month, there should be no problem in comparing the same periods between months. Note that this is no different from the way in which other products are dealt with, where prices can vary depending on the time of consumption, either in a fixed pattern (e.g. electricity, telecoms) or according to a more flexible pattern (e.g. flights, package holidays, hotels).

The longest interval of time scanner data can refer to is a month. In principle, as many days as possible should be included[10], but none should be included that refer to other months. It is important to ensure that the **time interval** is defined in the same way throughout the year.

Most commonly, scanner data is collected weekly, i.e. all transactions taking place during a week are aggregated. The current practice to calculate the unit value item code is to simply divide the total turnover for that item code for the period by the total quantities sold over the same period. The concept of unit values requires homogeneity of items.

A further consideration is how the delivery of the data fits into the HICP production cycle, especially when office staff have to make replacements and quality adjustments. Processing data per week at weekly intervals is granular enough to monitor developments in prices and check whether new item codes are relaunches or replacements; one does not have to wait until

[9] As there are many possible cases, no general guidance is given and practical solutions will have to be found case by case.

[10] See: A newly identified source of potential CPI bias: weekly versus monthly unit value price indices, Diewert, W.E., Fox, K.J and de Haan, J, Economic Letters, volume 141, April 2016, pp. 169-172.

all the weekly scanner data files have been delivered for a particular month before checking for replacements. Collecting monthly scanner data is impractical to fit into the production process, given the time constraints on that process.

4. Item codes

This chapter presents the various types of codes that can be used and addresses some general points regarding the sample of codes.

GTIN[11] is the name currently used for the code formerly known as EAN, which is the code most commonly used when dealing with scanner data. In addition to GTIN, the following codes may also be used: price look-up (PLU), in-store and stock-keeping units (SKUs).

The PLU codes are short and used by cashiers or customers to rapidly enter a code for an item in a cash register or another system that prints a sticker bearing a code that can then be scanned. The PLU code in Figure 4 is the number 3112, which should always refer to a Caribbean Red Papaya, regardless of the producer. Note that no part of the GTIN 7898921976015 corresponds to the PLU code.

Figure 4: GTIN and PLU

Figure 5: In-store code

In-store codes are GTINs with a prefix between 20 and 29 and are only valid for a given retailer. They refer to items that are given a code in the outlet. After weighing an apple as shown in Figure 5 an in-store code is printed, the last part of which encodes the price to pay: €0.81. The last digit is a checksum. The part of interest to the price statistician is the first part (2306803), which can be found in the scanner data file. For these codes, retailers may need to provide additional information so that a price per quantity can be calculated. Alternatively, the weight or quantity can be encoded in the last part and the price calculated at the check-out. It is therefore important to understand, for each retailer, what exactly the data refers to.

Besides GTINs, some retailers may use stock-keeping units (SKUs). These codes can be slightly more generic than GTINs. See Table 1 for an example of a product that does not change for the consumer and therefore has one SKU but two GTINs. A possible reason might be that the product is produced in two separate factories, and the producer wishes to

[11] See http://www.gtin.info for more information. The GS1 General Specifications are also recommended reading.

distinguish between them. Starting from week 39, the same SKU is sold under multiple GTIN codes for a few weeks.

Table 1: SKU and GTIN codes

Week	SKU	Product description	Unit (in grams)	Units sold (number)	Turnover (in euro)	GTIN
37	1234567890	Chocolate Brand x	0.375	380	2755	#8000565755675
38	1234567890	Chocolate Brand x	0.375	561	3540	#8000565755675
39	1234567890	Chocolate Brand x	0.375	1289	7657	#8000565755675 #8000508890089
40	1234567890	Chocolate Brand x	0.375	763	4288	#8000565755675 #8000508890089
41	1234567890	Chocolate Brand x	0.375	1128	6757	#8000565755675 #8000508890089
42	1234567890	Chocolate Brand x	0.375	912	5591	#8000565755675 #8000508890089
43	1234567890	Chocolate Brand x	0.375	621	4229	#8000565755675 #8000508890089

Which codes are used by the NSI will depend on the specific way in which the retailers' operations are organised, and on what data they are able and willing to supply. In all cases, the codes should:

1. Identify a unique product. Items with the same GTIN are identical, but if other codes are used it is important to make sure that the same code refers to the same physical item. If an in-store code is used to designate *any* bunch of flowers, then scanner data for that code is not useful for measuring price development, because the bunches of flowers would be incomparable.

2. Consistently refer to the same product over time. The lead-time to reusing a GTIN for an entirely different item is 48 months, except for clothes, where it is 30 months. This means that 48 months after the producer sold the last item with a particular item code, that code may be used for a different item.

3. It is likely that retailers will supply a mix of different types of item codes.

5. Obtaining scanner data

After determining the requirements in terms of regions, outlets and time an NSI can approach retailers with a wish list for obtaining scanner data. Whether or not the NSI's wishes can be met or a compromise found acceptable is something not addressed by this guide, save for three remarks:

1. Obtaining scanner data depends on the legal and institutional arrangements in each EU country and on the relationship between the NSI and retailers.

2. Getting the cooperation of a retailer to provide scanner data may be a lengthy process. A relationship of trust with a retailer should be established and nurtured.

3. The NSIs' wishes will develop over time, as will the availability of data.

The following seven recommendations form the basis for obtaining scanner data, each followed by a brief justification.

1. If possible, collect scanner data directly from the retailer.

   The data should be provided by the retailer, as it concerns their economic activities, on which they are legally obliged to report. If third parties like regulators or market researchers[12] are involved in the delivery of scanner data, this should be seen as a service to the NSI and the retailer, and it should be clear what processing steps the third parties undertake. The responsibility for ensuring that the relevant price indices are correct rests with the NSI, not the third party.

2. It is recommended that data be collected at item code level.

   Per item code: turnover and quantities sold, from which an average transaction price (unit value) can be derived; the unit of quantity; content of the package; tax rates; and further information that identifies the item, such as a brand name and a product description; if available, the code of a retailer-specific classification . A retailer-specific classification is a classification owned, maintained and used by the retailer.

   The data should contain a reference to the period to which it refers.

   If the retailer uses other codes, such as SKUs, these should be considered as an item's primary identifier, because they are more stable than a GTIN. SKU are retailer-specific as well and use a more general description. Nevertheless, it is recommended that the GTINs should still be supplied so that the composition of SKU can be checked if necessary.

   Information on the quantity (pieces, kg, litres etc.) is important for quality adjustments and for the HICP at constant tax rates (HICP-CT).

   It is essential to include information that identifies the product. The more information is provided, the easier it is to identify items and thus classify the item codes and identify replacements. This is the reason why including a retailer-specific classification is very useful, especially if this can be linked to ECOICOP. Information that could link temporarily discounted items that have a different item code from their regular counterparts is useful; often a SKU is used for this purpose.

3. It is recommended that scanner data be collected and aggregated each day, or at most over a week, i.e. total turnover and quantities sold per week. The period to which the data refers should be clearly indicated.

   Using weeks – and receiving data by week – will allow a NSI to use the first three full weeks of the month for the HICP, while the fourth week will often include days of the next month. Similarly, the first few days of the month may often be included in the scanner data from the last week of the previous month. If scanner data can be given

[12] Unless the NSI's solution is to outsource or subcontract part of the work or even the actual provision of data.

per day, a more fine-grained approach can be taken that would allow the inclusion of days where the week is split over different months.

Receiving data at least per week also adds an element of safety. If data cannot be delivered for a particular week, the index could still be calculated with the other weeks.

Collecting data for every week may be easier for the retail chain because the delivery follows a fixed rhythm.

4. It is recommended that scanner data be collected or aggregated over outlets that are homogeneous in terms of the service offered (often a certain concept or retail formula) and by region depending on national circumstances or requirements. For larger countries where items and pricing policy may differ between regions, it is especially important to collect scanner data per region.

   Retailers may have different types of outlet (concepts or formulas), such as large supermarkets, small supermarkets, mini markets in town centres and internet shops, all offering different levels of service. These different levels may have different pricing policies, and the numbers of shops of each type may change as well.

   If pricing policy is determined by region, a regional aggregation may be included, especially if scanner data has to be integrated with traditional price collection (from bakers, butchers and other traditional retailers.)[13]. For these reasons, it is recommended that data be collected by outlet. This data could also be aggregated to geographic regions by the retail chain, depending on the specific situation, but the different outlet types should be kept separate.

   If data storage is an issue, a sample of outlets could be considered.

5. The scanner data may contain transactions that should, as a matter of principle, be excluded from the HICP and these need to be discussed before introducing scanner data. The points relate to business transactions and returns[14].

   Scanner data may contain data on transactions between the retailer and other businesses. The HICP should not cover such transactions. How important these transactions are and if there is any effect on consumer price levels or development needs to be checked with the retailer. If necessary, such transactions should be excluded.

   Scanner data may include purchases of items which are returned within a given period after the purchase. This needs to be discussed with the retailer, especially how important they are, in which week they are recorded and thus how they influence price development, and whether they can be eliminated. In the case of food returns are probably not of importance but in the case of clothing they may well be important.

[13] Note that regional aggregation could also allow for aggregation over outlets of different retail chains within a defined region and that this could take outlet-substitution into account.

[14] These points is also relevant if scanner data is used for weighting

The importance of the issues, especially the first two, needs to be kept in perspective: how important are they really for price development?

6. Scanner data should, in line with HICP principles, include purchases of items sold at discount prices. It should be very clear which discounts are included and how. In practical terms, it may be difficult to filter out, for instance, discounts on food items approaching their expiry date. A related issue that needs to be clarified with the retailer is how coupons are treated in the data files.

7. The provision of scanner data should preferably be automated in a secure manner. This refers to both the extraction process at the retailer and the transmission of the data set. The delivery process should be automated, as this simplifies the delivery of the data to the NSI and thereby reduces the risk of errors.

8. The details of the provision of scanner data should preferably be laid down in a formal agreement. The HICP is an important statistic and the delivery of important input data should not be left to verbal agreements. The data is also highly confidential, and retailers will want to have guarantees on the confidential treatment of their data and the uses to which it is put. The annex contains three sample contracts[15].

9. A quality framework for scanner data should be used. The use of a quality report for scanner data is strongly recommended as it explicitly and systematically assesses data quality against requirements. Beyond this primary goal, it could be useful for the following purposes:

    - as a checklist when collecting scanner data: what topics to discuss with the retail chain or retailer;
    - as a list of issues to be dealt with in a formal arrangement with the retail chain;
    - as documentation or meta-documentation for internal users;
    - as a tool to monitor the quality of the data deliveries (checklists for data deliveries);
    - as documentation for satisfying requests on HICP compliance; and
    - as a part of an overall quality framework/programme.

    The current ESS quality report focuses on the output of statistics. At the moment, Eurostat[16] has no quality framework or report dedicated to input data, let alone scanner data.

    Table 2 provides an example of a possible quality report.

    As regards the general structure of the quality report, we have split the framework into the following three dimensions, as proposed by Daas in Secondary Data Collection:

[15] It should be borne in mind that the agreements discussed here lay down the details of the data delivery, not the legal requirement for the retailer to report on its activities.

[16] Work is currently under way to broaden the ESS quality framework and reports to include input data and the statistical processes. Please see also: United Nations Economic Commission for Europe - UNECE (2011): Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices. New York and Geneva

source, metadata and data. The *source dimension* focuses on the supplier and the delivery of the data. The *metadata dimension* contains indicators on how fit the data is for statistical use. The *final dimension* concerns the data itself. Together, these three dimensions cover all quality aspects relevant to scanner data.

Not all the indicators can be quantitatively measured, especially in the first two dimensions. In the third dimension, however, it will become more relevant to compare summary statistics on the data received with quantitative criteria and to monitor the development of the summary statistics over time. It can be difficult to set up such criteria, yet it is important to have a checklist against which to measure (and decide) when the data cannot be used.

It should also be clear that criteria develop over time and are dependent on the context. This is why the criteria presented below are just an example.

Table 2: Quality report for scanner data

1 Source		
This first set of quality indicators apply to the source of the data and address some general aspects of the relation with the retailer.		
DIMENSION	**INDICATOR**	**CRITERIA**
1.1. Retail chain	Contact	▪ Name of retail chain/data source ▪ Retail chain contact information
1.2. Relevance	*Usefulness*	▪ *Importance of source for NSI (market share e.g.)*
1.3. Security	Security	▪ Manner in which the data is securely sent to NSI
1.4. Delivery	4.1 Agreements	▪ Are the terms of delivery documented?- Frequency of deliveries
	4.2 Punctuality	▪ How punctual can the data be delivered? ▪ Time-lag with which exceptions are reported
	4.3 Format	▪ Format in which the data is be delivered
	4.4 Service in return	▪ Details on any service provided in return
1.5. Procedures	5.1 Planned changes	▪ Familiarity with planned changes of data source ▪ Ways to communicate changes to NSI ▪ How long before the change will the NSI be informed
	5.2 Fall-back scenario	▪ Emergency measures when data source is not delivered according to arrangements made

2 Metadata		
The metadata quality indicators indicate how clearly the coverage is defined and how these compare to the NSI's requirements, the identification of products and the degree to which the data has been checked and modified by the retailer. Most of these indicators will be discussed with the retailer in an attempt to reconcile the wishes of the NSI with the data that the retailer can easily supply. It is essential to understand how scanner data has been put together and what is included and excluded, what data editing the retailer does.		
DIMENSION	**INDICATOR**	**CRITERIA**
2.1. Clarity	1.1 Delimitation of the retailer	▪ Is it clear to exactly which parts (outlets, divisions) of the retail chain the data refers?

Table 2: contd.

	1.2 Types of transaction	▪ Clarity of the types of transactions included. ▪ Is all consumer-related turnover included in the data? ▪ Is all business related turnover excluded? ▪ Are returns and refunds included? All? ▪ Are discounts included? Which?
	1.3 Turnover definition	▪ Is the definition of turnover clear? ▪ Includes VAT? Include returns, and if so, how? ▪ In the week they were originally bought or in which they were returned? ▪ What is the format?
	1.4 Quantity definition	▪ Is the definition clear? Do the quantities include returns or not?
	1.5 Unit of quantity definition	▪ Is the definition clear? What is the format?
	1.6 Periodicity	▪ Clarity of the period to which the reported data relates.
2.2. Data treatment	Checks (by retailer)	▪ Population unit checks performed ▪ Variable checks performed ▪ Combinations of variables checked ▪ Extreme value checks performed

3 Data		
The quality indicators that concern the data actually delivered and should be monitored with each delivery.		
DIMENSION	**INDICATOR**	**CRITERIA**
3.1. Technical checks	1.1 Readability	▪ Accessibility of the file and data in the file
	1.2 File declaration compliance variables	▪ Compliance of the data in the file to the metadata agreements
	1.3 Convertibility	▪ Conversion of the file to the NSI-standard ▪ % of records that could not be converted to the NSI standard
3.2. Accuracy	2.1 Inconsistent objects	▪ Extent of erroneous objects in source ▪ % of GTIN codes with illogical relations to (aggregates of) objects. For example: % of GTIN codes with missing values for variables, % GTINs with incorrect VAT rates and % of GTINs with illogical prices
3.3. Complete-ness	3.1 Coverage	▪ Absence of target GTIN (missing objects) in the source ▪ Check on the number of GTIN codes per category in the shop classification against a – historically determined – expectation. ▪ Check on the total level of expenditure against a – historically determined expectation. ▪ Presence of non-target objects in the source
	3.2 Dynamics of objects	▪ Changes in the population of objects (new and dead objects) over time
3.4. Time-related	4.1 Timeliness	▪ Time between the end of the reference period and receipt of the source
	4.2 Punctuality	▪ Time lag between the actual and agreed delivery date

6. Processing scanner data: overview

In general, there are three steps in processing scanner data and two approaches to determining the sample of items. The three steps are:

1. Classifying item codes to ECOICOP (Chapter 7).
2. Sampling of item codes and the calculation of the elementary aggregate index (Chapter 8), and
3. Integrating the outcomes from the scanner data production process into the HICP (chapter 9).

The two approaches to sampling items (step 2) for supermarkets are the ***static*** and the ***dynamic*** approach.

Assume that the data set is complete and correct, classifying item codes to ECOICOP is the first step in processing scanner data. The ECOICOP aggregate may be a nationally defined aggregate below the level ECOICOP subclass level.

The classifying of item codes is followed by the sampling of item codes. In the static approach, a sample is drawn from year t and used for 12 months following December of year t. The sample is kept and replacements are made as needed. The dynamic approach draws a matched sample of items over a period of two months (t and t-1), moving up each month.

After the sampling of item codes has taken place, average prices or indices are calculated as necessary and integrated into the HICP. The remainder of this guideline will treat these topics in greater detail.

The ***static approach*** closely mimics the traditional fixed sample including the treatment of relaunches. Towards the end of a year a sample of item codes is purposively drawn as representatives for the following year. If an item becomes less representative or disappears during the year, it is replaced. This all follows traditional methodology, but with the advantage of having full information on actual transactions on which to base choices about the initial selection of item codes and, if needed, their replacement during the year.

This method has advantages, especially if scanner data is used on a limited scale and if it has to be combined with data collected in the traditional way. In such cases, it may be convenient and efficient to 'hand-pick' those item codes that best fit the product descriptions used in traditional price collection. The method also has drawbacks; it is labour-intensive and makes limited use of the available data.

The ***dynamic method*** automatically selects a representative sample of item codes for each consecutive set of two months (t and t+1, t+1 and t+2, t+2 and t+3 and so on) by selecting all matched item codes that have a turnover above a certain threshold and will include new and sufficiently important items whilst dropping items that are less important. The method resembles monthly replenishment and chaining. The dynamic method is favoured when substantial amounts of scanner data have to be processed because it can easily be automated. However, if relaunches and replacements occur frequently, they need to be dealt with separately to ensure quality adjustments are made.

When processing large quantities of data, mistakes are bound to occur and should be seen as acceptable within certain limits. It is therefore important that procedures are put in place to check the quality of the automated procedures, monitor the number and nature of mistakes and errors, and ensure that the automated systems are improved. Automatic processing scanner data requires at least two quality checks to be made:

1. Checks on whether items have been correctly mapped to ECOICOP. This is valid for both the static and the dynamic approach.

2. Checks on whether replacements have been included correctly. This is particularly relevant for the dynamic approach.

The outcomes of processing scanner data, indices or prices, are of course subject to the regular validation and plausibility checks used for the whole HICP.

7. Classification

Classifying item codes to ECOICOP is a new aspect or step in the production of the HICP. Given the large numbers of item codes involved, manually classifying codes to ECOICOP is not advisable. Modern techniques taken from semantic technologies, artificial intelligence and machine learning are more appropriate. Eurostat has started a separate project with external partners to develop a classifying methodology and this section will be extensively updated in 2018.

The first step in processing scanner data is classifying the items to ECOICOP. The level of product classification to which the item code is mapped will be the lowest level of classification used by the NSI; often the 6- or 7-digit level of ECOICOP.

Classification is done in two steps:

1. The initialisation of a new retailer, when all the data needs to be mapped.

2. Classifying as part of the monthly production cycle.

Classifying item codes is partly unique to each retailer, as the data each of them provides differs and descriptions for the same item code need not be the same across all retailers.

At the initialisation stage, it is important to understand which item codes should be used and how stable the item codes are, especially if PLU, in-store and SKU codes are used. The scanner data set may well include groups of items, such as clothing or home decoration, that are to be excluded or assigned to a different division of ECOICOP. If item codes have unclear meanings, then they are not useful and should be excluded, or alternatives must be discussed with the retailer.

Given the large number of item codes, manually classifying each item code is not realistic in terms of the resources needed, and automated procedures are preferred. Often a retailer-specific classification can be used as a shortcut. It is also possible to simply replace the 6- or 7-digit level with the detailed retailer-specific classification aggregates. For example, before the introduction of scanner data one NSI defined some 175 elementary aggregates in

ECOICOP division 01 and with scanner data now defines 400 elementary aggregates based on the retailer-specific classification (see Chapter 9.3).

The classifying procedure, especially if parts of it are fully automated, may lead to an incorrect classification of a few item codes to ECOICOP. Building procedures that are 100% water-tight is not a practicable solution to this, as the cost would be prohibitive. It is better to monitor the automated systems separately from the production cycle, using a sample of newly mapped item codes, accept that some mistakes are likely, and use the results to improve the automated classification procedures. The number of mistakes should, of course, be below a certain threshold. If the threshold is exceeded, corrective action may be called for before using the outcomes to calculate the HICP.

7.1. Initialisation of scanner data from a new retailer

The initialisation consists of an in-depth study to understand the churn of item codes, identifying items and groups of items that are to be excluded and developing automated classification routines. For each retailer a separate initialisation is required.

a) In principle, the classification process should be automated as much as possible. The use of automated tools is recommended for the semantic analysis of item descriptions to classify items.

b) The type of codes (GTIN, PLU, SKU, etc.), the descriptions (the meanings of abbreviations, etc.) and other metadata linked to the codes must be fully understood.

c) The churn of the item codes must be understood. To achieve this, a longer period has to be studied. Do codes refer to the same item each month? What is the attrition rate[17] and the rate at which new item codes are introduced? The results of these tests will determine whether a replacement strategy needs to be implemented, and if so, how.

d) Use of the retailer-specific classification is recommended. This should be mapped onto the lowest level of the national ECOICOP used. A retailer-specific classification can be seen as a short-cut; if the retailer classifies an item as 'white rice', then one may assume that all items with that classification should be classified under ECOICOP 01.1.1.1 (Rice).

e) The items or groups of items should be identified to exclude, for instance, item codes with unclear descriptions. For example, a code could be used to designate the 'special pastry offer' or the 'bunch of flowers' which changes from one day to the next.

7.2. Monthly classification process

f) The automated classification process should be used to map **new** item codes and item codes with changed meta-data (description, classification etc.).

g) For item codes that could not be mapped unambiguously, a classification algorithm should be used. If this is not possible, new item codes will have to be classified

[17] Attrition rates are the percentage of GTINs that disappear in the next period.

manually, by visual inspection of the item description, and preferably by focusing on item codes with a high turnover and item codes that are sensitive to changing item descriptions, such as those for fresh produce.

h) Each of the following checks should be carried out regularly:

   i. Monitor the changes in the retailer classification, such as item codes that move to another group.

   ii. Check the classification of item codes where the description has changed.

   iii. Check if item codes retain the same retailer-specific/ECOICOP classification.

   iv. Check item codes that have been excluded, such as items with an unclear description.

   v. Monitor new and disappearing item codes (this is part of the method of dealing with replacements).

i) Ensure that results can be reproduced: save previous versions of the retailer-specific classification and the classification to ECOICOP.

j) The quality of the classification, whether or not it is automated, needs to be checked. It is advisable to randomly select a sample of item codes and check the correctness of the classification. Errors should then lead to improvements in the classification procedures, which may or may not be automated. There are different ways to carry out the checking, such as a visual scan.

7.3. Classifying with a retailer-specific classification

After a thorough analysis of the retailer-specific classification, all the item codes of an aggregate that fall within or coincide with some lowest level ECOICOP can be directly linked to ECOICOP.

After this, an analysis has to be made of the retailer-specific classification aggregates that could not be mapped, such as 'Asian food', which could include rice, vegetables, condiments, and so on. If reassigning item codes would have a significant impact in terms of the turnover of the elementary aggregate to which the item codes properly belong, then item codes should be reassigned. If a retailer-specific classification aggregate is considered insignificant in terms of turnover, it may be left out altogether. However, it should still be monitored, so that if it exceeds a certain threshold the items are again included.

The classification should be repeated *each month* for all new item codes. During the monthly classifying, potential replacements could be filtered out by monitoring new item codes and those on the way out owing to decreasing turnover and quantities.

A further number of checks should be programmed that monitor changes in the retailer-specific classification, item codes that change classification, and item codes where the description changes.

If the whole retailer-specific classification changes, re-classifying of the classification to ECOICOP may be necessary. It is generally not in retailers' interest to change the

classification too often. However, the retailer should be made aware of the need to inform the NSI in good time if changes in the classification are planned.

7.4. Classifying without a retailer-specific classification

If no retailer-specific classification is available or it is not usable, alternative methods are needed. Classifying item codes manually is not recommended, as it would require more resources. A better option is to use classification algorithms.

8. Sampling: static and dynamic approach

The static and dynamic approaches were introduced in Chapter 6 and will be described in more detail in this chapter. First we provide some background, then a step-by-step description.

The main difference between the two methods is the automated monthly updating of the sample in the dynamic approach. The values for the filters used in the dynamic method could also be used in the static approach.

8.1. Static approach

The static approach closely mimics traditional price collection methods.

The first step is to make an initial sample of item codes in December after which, for each month, the sample is maintained as in a traditional survey. The selection over a longer preceding period is needed to ensure that the sample is representative for the whole year (January to December), thereby excluding the sales typical for December.

A balance must be struck between representativity and the ability to maintain the sample. The extent to which the sample can be maintained is influenced by several related factors: the size of the sample, the churn and the efficiency of the system in terms of suggesting replacements that are essentially equivalent to the items originally sampled. The filters used monthly in the dynamic approach are also useful for determining the sample in the static approach.

One NSI that has implemented this method makes an initial sample by selecting all those item codes that constitute 50% of turnover for the selected elementary aggregate (ECOICOP level 6, non-seasonal item) and that were sold throughout the 12 months of the base period. Each elementary aggregate has at least 5 item codes, and some more volatile elementary aggregate have a few more added. The resulting number of item codes is sufficient and maintainable. Another NSIs implementation contains some 6000 SKU codes of which 2–3% per month have warning tags indicating that 'something' needs to be checked manually.

Monthly cycle in the static approach

In the monthly cycle, a comparison is made between the price reference period and the current month, and replacements are made together with possible quality adjustments.

The use of scanner data, with the option of looking back into the sampling frame of many previous months, allows for a judicious replacement strategy. If scanner data is processed per week, changes in item codes can be monitored in good time. Consequently, imputations are rare because replacements can be introduced quickly.

Given the updated sample, the prices can be extracted and entered into the regular production process system, where they will be validated alongside the traditionally collected prices.

8.2. The static approach in steps

Annual updating of the sample

a) The sample of items for year t should be based on turnover from a sufficiently long period, if possible the whole year t-1, to ensure that the sample is stable. The period chosen should account for seasonal items. December t-1 is the price reference period.

b) It should be ensured that in the initial fixed sample the items cover a sufficiently high percentage of turnover in each elementary aggregate, and that the sample is relatively stable. Depending on the number of items and the distribution of turnover over the items, a coverage level of between 50% and 80% is not uncommon. Because the details differ across countries, retailers and elementary aggregates, it is impossible to provide any precise guidance.

c) The resulting fixed sample size should be sustainable in terms of the staffing resources needed to maintain it.

Monthly production process

d) The occurrence of items in the observation month, compared to the previous month must be checked every month. This should preferably be done each week, assuming the data is delivered on a weekly basis.

e) If an item code is not found, a replacement is sought within two months, in accordance with the standard HICP requirements regarding missing items.

f) To ensure that the sample is representative, items whose turnover (or number of sales) falls below a specific percentage or value of December t-1 should preferably be replaced.

   The precise percentage should be determined in a pragmatic way: a sufficiently large turnover should be covered and a sufficiently large number of item codes included, all of this relative to the total number of item codes, total turnover and churn. When replacements are made, quality adjustments should be made if necessary, and for supermarkets these are often limited to package size adjustments. If no replacement can be made, the item is discontinued until the annual resampling.

g) This method replaces an item code if, and only if, the code is no longer representative. An alternative is a method that includes all item codes above a given threshold during the observation period. This makes the method more dynamic, in that new items can be included in the index sooner.

h) If prices are the output of the scanner data production process, it is recommended that they be processed further in accordance with standard procedures for validation and checking plausibility of outcomes.

i) If indices are the output of the scanner data production process, they can be plugged into the aggregation schema (see Chapter 9.3.)

8.3. Dynamic approach

The dynamic approach was introduced to improve the quality of the HICP without a large increase in the resources associated with the manual labour involved in the static approach. The dynamic approach would also allow an NSI to increase the number of scanner data retailers or outlets.

The monthly process draws a sample of those item codes that are present in both the current month and the preceding month and that represent a large portion of turnover for that elementary aggregate. It is an empirical fact that often many items contribute to the broad selection of goods on offer, but relatively few are responsible for a large share of turnover. The dynamic method takes this fact into account, as we will see below.

The dynamic basket uses a set of filters and an algorithm to select a matched sample for each month and the preceding month. The entire sampling procedure can be fully automated. However, this convenience comes at a cost: the system does not necessarily link relaunches, because there is nothing inherent in the item code that links two item codes, unless SKUs are used. Hence relaunches have to be treated outside the system. This is also the reason why the dynamic method is not suitable for assortments with a high churn.

The filters should be developed in the initialisation phase and regularly monitored during the production period.

The system uses a blacklist, two filters and an algorithm to select a sample of matched item codes:

1. A blacklist that removes groups of codes (e.g. for clothing) or item codes that are unusable for some reason e.g. codes that refer to 'a bunch of flowers', where the exact composition of the flowers changes daily.

2. A dump filter that removes items if sharp falls in price and turnover suggest that the product will be taken off the market and cease to be representative. The aim is to eliminate the downward pressure of clearance prices on the index. This is needed (despite point 4 below) to ensure that items leaving the market are removed at the right time, because the turnover may, even in the second month, still be high enough to pass the low-sales filter.

3. An outlier filter which removes prices that drop below or rise above given thresholds. This filter ensures that clear errors, such as decimal errors, are removed, as are items where the discounts are so large that prices can drop practically to zero. An example is when consumers can use a coupon to receive a specific item free of charge. Such outliers should of course be investigated and the causes of the error remedied.

4. A low-sales filter that filters out item codes with very low sales, or, conversely, ensures that the selected codes represent a sufficiently high proportion of turnover (between 50 and 80%). The low-sales algorithm is defined as:

$$\frac{s_{m-1} + s_m}{2} > \frac{1}{(n \times \lambda)}$$

where s_{m-1} *and* s_m is the turnover share in month m–1 and m, *n* is the number of items in the elementary aggregate and λ = 1.25. The value 1.25 for λ is empirically determined so that the selected item codes represent about 80% of turnover. Since this method selects item codes on the basis of turnover, certain product segments like low-value or brandless items must not be included in the sample, as their low prices also contribute to a low turnover.

The exact values for the filters cannot be given, as they depend on the specifics of national markets and probably vary between countries.

Relaunches and replacements are a potential problem for this method because as the system does not automatically link a disappearing item code with its relaunch or replacement item code.

The imputation period should be set for 14 months to ensure that items (automatically) re-enter the computation system and that the elementary aggregate indices satisfy the identity test and the chained index passes the transitivity test. The period of 14 months is a little longer that 12 months and ensures the inclusion of item codes that are sold each year for a short and possibly shifting period, e.g. Easter items. The period of 14 months also ensures that if an item re-enters the calculation after that period it is treated as a new item and the risk of comparing item codes for two different items (re-use of an item code by the producer for another items) is minimal.

Items are not explicitly weighted. The turnover is only used for sampling items and calculating the unit value for each item.

Monthly production process

a) The filters developed as part of the initialisation process for scanner data are applied.

b) For relaunches and replacements, old and new item codes should be combined and unit value indices should then be calculated for the combined codes. Quality adjustments should be applied as needed, especially for changes in package size.

c) The elementary aggregate index is calculated on the basis of the matched set of representative item codes for items that are actually sold in two successive periods. An unweighted Jevons index is calculated over the current and preceding month as follows (see HICP Methodological Manual, formula 8.11)

$$P_J^{(m-1)t,mt} = \frac{\left(\prod_{k=1}^{K} p_k^{mt}\right)^{\frac{1}{K}}}{\left(\prod_{k=1}^{K} p_k^{(m-1)t}\right)^{\frac{1}{K}}} = \left(\prod_{k=1}^{K} \frac{p_k^{mt}}{p_k^{(m-1)t}}\right)^{\frac{1}{K}}$$

where K denotes the set of common item codes belonging to the elementary aggregate K.

Prices for item codes that are not present in subsequent periods are imputed[18] by the price development of the elementary aggregate for a period of 14 months to ensure that seasonal items re-enter the index at the correct time, allowing for shifts between years due to the weather and holidays such as Easter.

The chain-linked index is then as follows (see HICP Methodological Manual, formula 8.13):

$$CP_J^{0t,mt} = P_J^{0t,1t} \cdot P_J^{1t,2t} \cdot \ldots \cdot P_J^{(m-1)t,mt}$$
$$= \frac{\left(\prod_{k=1}^{K_1} p_k^{1t}\right)^{\frac{1}{K_1}}}{\left(\prod_{k=1}^{K_1} p_k^{0t}\right)^{\frac{1}{K_1}}} \cdot \ldots \cdot \frac{\left(\prod_{k=1}^{K_m} p_k^{mt}\right)^{\frac{1}{K_m}}}{\left(\prod_{k=1}^{K_m} p_k^{(m-1)t}\right)^{\frac{1}{K_m}}} \neq \frac{\left(\prod_{k=1}^{K} p_k^{mt}\right)^{\frac{1}{K}}}{\left(\prod_{k=1}^{K} p_k^{0t}\right)^{\frac{1}{K}}} = P_J^{0t,mt},$$

where K_1 denotes the set of common items in period 0 and 1, K_2 the set in period 1 and 2 and so on. Since new items appear and others disappear, the chain-linked month-on-month index does not reduce to the direct Jevons index. It is thus important to investigate the degree to which the sets $K_1, K_2 \ldots K$ differ, so that the NSI can assess the risks of downward drift. If the changes are substantial, as may be expected for clothing or cosmetics, this method should not be used[19].

Superlative indices (Fisher, Törnqvist) may not be used in combination with chain-linking, as these formulas may lead to considerable drift in the index[20].

d) A quality control system that is independent of the production system should be used It should check regularly whether or not the production system handles replacements in a correct manner.

There are risks involved in using highly automated systems for the monthly production in which filters and algorithms make the choices. The outcomes (price indices at some level of aggregation) will need to be checked for plausibility using the normal procedures.

To offset these risks, the performance of the algorithms needs to be checked regularly and independently, as these algorithms assume a degree of stability in the population on which they operate. There may come a time when this assumption is no longer valid.

[18] There is some tension between this recommendation and existing HICP legislation that needs to be resolved in the light of the use of scanner data.

[19] For an example of a situation in which this method can clearly not be used, see the dresses in Fig. 3 in Chessa, T, A new methodology for processing scanner data in the Dutch CPI, Eurona, Volume 1/2016, pp. 50-71.

[20] See, for instance, Ivancic, L., Diewert, W.E. and Fox, K.J., Scanner data, time aggregation and the construction of price indexes, Journal of Econometrics, 2011, Volume 161, pp. 24-35.

9. Seasonality, HICP-CT and Integration

Before discussing the integration of scanner data into the overall HICP aggregation, we will examine the HICP CT and seasonal items.

9.1. Seasonality

Seasonal items are those that are not available throughout the year (*strong seasonality*) following some cyclical pattern. In the case of *weak seasonality* sales take place in all periods but vary between periods.

Strong seasonality is dealt with in HICP regulations, and scanner data for seasonal items can be dealt with in accordance with those regulations. However, scanner data offers the possibility of taking a more fine-grained approach, as it contains data that is more detailed and up-to-date. Scanner data makes strong seasonality in expenditure shares and prices explicit, whereas traditional data sources may not do so at the same level of detail.

Scanner data also allows the start and end of the seasons to be monitored more precisely. This means the system should do two things. Firstly allow a long enough period of imputation so that the item codes can be compared between the old and new seasons. Secondly accommodate for shifts in seasonal patterns. The imputation period is therefore set at 14 months. This ensures that items (automatically) re-enter the computation system and that the elementary aggregate indices satisfy the identity test and the chained index passes the transitivity test. The period of 14 months ensures the inclusion of item codes that are sold each year for a short and possibly shifting period, e.g. Easter items.

9.2. HICP CT

For the HICP-CT, the prices should be estimated on the basis of the metadata available for each item code. If no such data is available, the calculation may be based on an estimate of the average tax rate applicable to the elementary aggregate. If this is the solution chosen, the elementary aggregate should be chosen in such a way as to ensure that the item codes are homogeneous in terms of the tax regime applicable. This means that item codes with and without excise duty should not be mixed.

The HICP-CT indices can be calculated at the level of detailed elementary aggregate indices. This would be the preferred solution in a situation where large quantities of prices are processed.

9.3. Integrating scanner data in the HICP

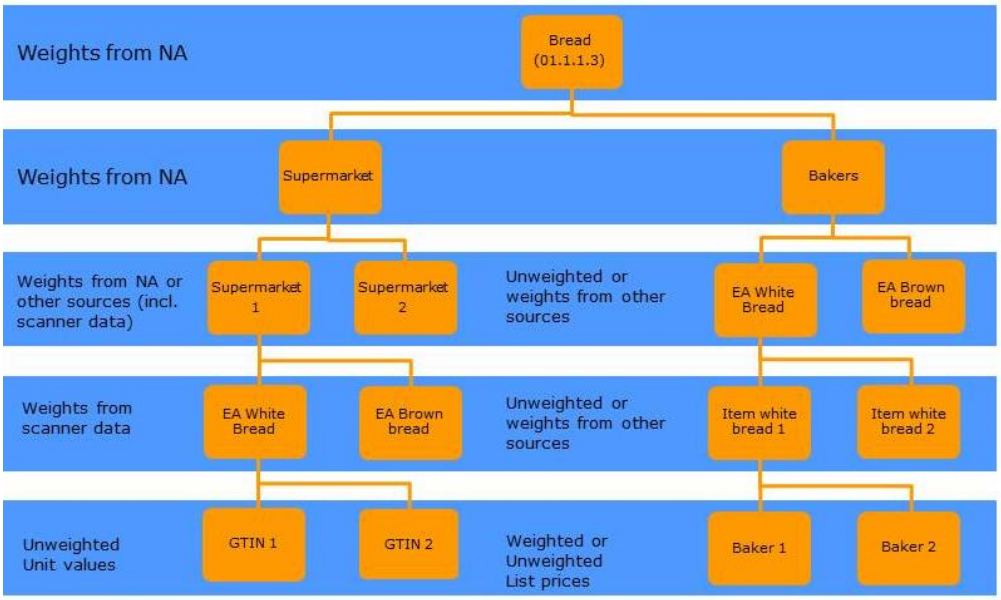
Scanner data has to be integrated with traditionally collected prices simply because not all price data is collected with scanner data. At some stage in the aggregation scanner data has to be combined with traditionally collected data.

Two different cases can be distinguished:

1. where scanner data covers a part of consumer expenditure for a specific ECOICOP aggregate;
2. where scanner data from supermarkets covers all of consumer expenditure for a specific aggregate.

In the first case, the aggregate weight cannot come from scanner data only and traditional sources must be used. For example, see Figure 6 below, if bread is sold by supermarkets (scanner data) and bakeries (traditional price collection), then the total weight for bread would come from National Accounts. The weight for bread could then be split between bakeries and supermarkets. For supermarkets, weights can be further broken down using scanner data.

Figure 6: Integrating scanner data and traditional data



Note that a supermarket for which no scanner data is received can be treated as a baker. A second thing to note is that if this approach is taken, all bakery items from a supermarket could potentially enter the index, whereas for the traditional bakeries only the items selected for price collection enter the index. Finally for the supermarkets the elementary aggregates could be different for each supermarket. For example, if one supermarket would classify all the bakery products into white-, brown and other bread then one could use these three elementary aggregates. If another supermarket would distinguish between fresh and pre-packed bread, then these two elementary aggregates could be used. What limits the freedom to choose elementary aggregates is the level of publication; if one where to publish brown- and white bread then the items of the second supermarket would have to be reclassified.

An example of the second case can be given by simply leaving out the bakeries in the example given above.

After prices or indices have been integrated with other parts of the production system, the results should be subjected to the same validation and plausibility checks as the other parts.

Annex 1 Examples of contracts

CONFIDENTIALITY AGREEMENT

Agreement number: 2013/ H

BETWEEN

THE BELGIAN STATE (THEMATIC DEPARTMENT PRICES OF STATISTICS BELGIUM OF THE FPS ECONOMY, SMES, SELF-EMPLOYED AND ENERGY)

AND

THE SUPERMARKET CHAIN

The Belgian State (Thematic Department Prices of Statistics Belgium of the FPS Economy, SMEs, Self-Employed and Energy), with registered office in *xxx*, Enterprise number: 0314.595.348, hereinafter called 'Statistics Belgium', for the purpose of this agreement represented by xxx, Deputy Prime Minister and Minister of Economy.

And

The supermarket chain, address, for the purpose of this agreement represented by *XX*

Observe that:

- In order to calculate the consumer price index, as well as to meet its European obligations with regard to the harmonised index of consumer prices, the project average prices and purchasing power parities, Statistics Belgium is required to carry out price observations in shops that are part of *the supermarket chain*;
- Statistics Belgium tries to use as many electronic files on sales and turnover at product level as possible for the statistics mentioned above,
- in order to further improve the reliability of these statistics;

Have agreed:

Article 1. Transmission of data to Statistics Belgium

1.1. *The supermarket chain* shall transmit the data only to Statistics Belgium.

1.2. *The supermarket chain* shall transmit data to Statistics Belgium in electronic form each week for the production of the statistics mentioned above. The transmitted data shall at least be broken down by week, i.e. from Monday to Sunday.

1.3. To ensure the timely production of these statistics, the data mentioned in article 2 shall be transmitted at the latest on the second working day of the week following the week to which the data refer. A working day is understood to be every day of the week, excluding Sundays and bank holidays.

1.4. Data shall retroactively be transmitted from 1 January 2012, these historical data shall at least be broken down by month. Data shall be transmitted through a secured connection.

Article 2. Content of the transmitted data

2.1. The data to be transmitted to Statistics Belgium are defined in annex 1 of this agreement.

2.2. If *the supermarket chain* makes a selection of the sold products in the data, these will be provided for all products from its range that can be traced to the product groups in annex 2.

2.3. The data for each product may be aggregated for all points of sale of *the supermarket chain* combined. All points of sale are understood to be those for which *the supermarket chain* can provide data.

2.4. If *the supermarket chain* comprises multiple retail formulas, the data shall be broken down by retail formula wherever possible. For the application of this agreement, *the supermarket chain* shall provide data for the following retailers: ….

2.5. Reporting comprises both goods that are bought directly at the shop as well as distance selling.

2.6. The provided data only refer to sales to consumers.

Article 3. Use of the data by Statistics Belgium

3.1. Statistics Belgium shall use the data exclusively and only (1) for the calculation of the consumer price index, as well as to meet its European obligations with regard to the harmonised index of consumer prices, the project average prices and purchasing power parities and (2) to deliver a market report to Comeos in which aggregated volume and price data for product groups as listed in annex 2 are provided.

3.2. Based on the data, Statistics Belgium shall produce aggregated statistics to fulfil its task of providing official information for public use. To succeed in this task as well as in related tasks, it shall meet the provisions included in the law of 4 July 1962 on public statistics.

3.3. In accordance with the law of 4 July 1962 on public statistics, Statistics Belgium shall ensure that no information related to the individual situation of *the supermarket chain* can be derived from the data published by the Thematic Department Prices of Statistics Belgium.

3.4. If the delivered data fall within the scope of article 111 of the NAI law containing social and various provisions (21 December 1994), the agreement shall be legally void.

Aricle

Article 4. Transmission of the data to third parties

4.1. The supermarket chain shall transmit the individual data referred to in article 1 exclusively to Statistics Belgium.

4.2. Statistics Belgium may not transmit the individual data to a third party.

4.3. The supermarket chain shall give its permission to use the provided data for an aggregated report made available to Comeos (not-for-profit organisation).

4.4. The report may not be disseminated as an official statistic, since it is not representative of the entire market.

4.5. Statistics Belgium shall carry no responsibility for the use of the report by Comeos. In case of disputes related to the use of this report, the supermarket chain shall turn to Comeos.

Article 5. Unforeseen circumstances

5.1. If - due to unforeseen circumstances - the electronic data cannot be transmitted in time, *the supermarket chain* shall inform Statistics Belgium of the reason and expected duration of the delay one working day after the period referred to in article 1.2.

Article 6. Revision

6.1. Each party to this agreement may request its revision.

6.2. All changes to this agreement shall occur by means of a written annex, which will be agreed on under the same conditions as this agreement. A verbal agreement between parties is not binding.

Article 7. Duration and termination of this agreement

7.1. This agreement shall go into effect on *1 September 2013* and shall be subject to a trial period of 3 months. After this trial period the agreement will automatically be renewed each year, unless one of both parties in writing asks for the termination of the agreement, at least 3 months before the end of the duration of the ongoing year.

7.2. The first year referred to in article 7.1 runs from the end of the trial period to 31 December 2014.

7.3. Each shortcoming by Statistics Belgium with regard to the obligations in this agreement grants *the supermarket chain* the right to end the agreement with immediate effect, after Statistics Belgium has been delivered a notice of default by registered letter to remedy the shortcomings, and if this notice of default has not yielded an adequate outcome at the latest 8 working days after its sending.

Article 8. Applicable law and competent courts

8.1. This agreement shall be governed only by Belgian law.

8.2. If, due to circumstances or occurrences, problems arise with the execution or interpretation of this agreement, the parties shall commit to trying to reach an agreement that meets the demands of both parties and that is in accordance with the law of 4 July 1962 on public statistics and its implementing decisions, before any other steps are taken.

8.3. Any dispute arising from this agreement shall be subject to the exclusive jurisdiction of the courts of Brussels.

Done in Brussels on (date) in as many original copies as there are parties to the agreement. Each party acknowledges to have received an original copy.

For the supermarket chain	***For* the Belgian State (Thematic Department Prices of Statistics Belgium of the SPF Economy, SMEs, Self-Employed and Energy)**
Name	***Name,***
Function	***Deputy Prime Minister and Minister of Economy,***

List of annexes:

1. *List of data to be provided per product*
2. *List of the products concerned*

Contract

Supplier, for the purpose of the present contract represented by *name / position*

and

Statistics Netherlands, in this document further referred to as CBS, for the purpose of the present contract represented by *name / position*

taking into consideration that:

- for the calculation of the consumer price index (CPI), CBS is obliged to conduct price observations at branches of *supplier*;
- CBS aims to conduct price observations and observations on behalf of as many statistics as possible, such as turnover statistics, by using available electronic data on sales and volumes of items;
- The above will improve the reliability of the consumer price index;

have agreed as follows:

Article 1 Provision of information to CBS

1. *Supplier* shall provide weekly data in electronic format for the monthly production of the CBS consumer price index.
2. *Supplier* shall provide the data in accordance with the agreed record description as specified in Annex 1 to this document.
3. For timely processing of the monthly CPI, data covering one week (Monday up to and including Sunday) shall be in possession of CBS on the following Tuesday before 12.00 p.m.
4. The data shall be transmitted via a secure line.

Article 2 Use by CBS

1. CBS shall use the data referred to in Article 1 solely for statistical purposes.
2. CBS shall use the data referred to in Article 1 to fulfil its mission to provide official information for public use, and in performing this task or related resulting tasks, shall comply with all provisions under the Statistics Netherlands Act (*Staatsblad 2003, 516*).
3. CBS shall ensure that no data on prices of individual supermarket chains shall be able to be deduced in any way from data published by CBS. This is in accordance with Article 37 of the Statistics Netherlands Act.
4. CBS shall not provide data on individual firms, households or institutions related to third parties under any circumstances, unless this is explicitly communicated and approved by both parties (see Articles 5 and 6).

Article 3 Data delivery free of charge

Given the public role of CBS, the datasets shall be supplied free of charge.

Article 4 European Comparison Programme

The data shall be used for the European Comparison Programme, a mandatory EU statistical requirement enabling the determination of differences in price levels between countries. If the data are inadequate for this purpose, additional price observations shall be conducted at the branches. This will be communicated in advance with *supplier*.

Article 5 Delivery to third parties

The data may not be provided to a third party, i.e. a party other than the two parties to this contract.

Article 6 Unforeseen circumstances

If - as a result of unforeseen circumstances - electronic data on prices and turnover are not available on time, *supplier* shall inform CBS as soon as possible about the nature and duration of the malfunction. In such cases CBS shall be permitted to conduct substitute price observations in two branches.

Article 7 Reciprocity

1. The subsequent delivery of information from CBS to *Supplier* shall take place on the Friday following publication of CBS' press release (usually on the first or second Thursday of the month):

   - Price indices and Classification of Individual Consumption by Purpose (COICOP) for the chain of *supplier* and total (standard publication level). For all COICOP groups in which the prices of *supplier's* chain are included in the calculation,

   - The price developments of *supplier's* chain at a more detailed level,

2. The data shall be transmitted via a secure line.

3. This reciprocal delivery shall be free of charge.

Article 8 Duration of this contract

1. This contract shall come into effect on ...[*date*]… and shall remain valid for the period of one year. After this period the contract shall be automatically extended by one year, unless either party requests termination in writing at least three months before the end of the term of the current year.

2. Changes during the duration of the agreement may only be introduced if agreed, recorded in writing and signed by both parties.

3. Following annulment, suspension or termination of the contract, the provisions in the Statistics Netherlands Act remain valid.

Agreed and signed in duplicate,

The Hague, …[*date*]…

Supplier	CBS
Position	Manager Department of price statistics and short-term indicators
Name	*Name*

Annex 1: Record deliveries Description