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

Draft

May 2017

Foreword

Scanner data is a relatively new source of data for producing consumer price statistics. Currently a fifth of the Member States use scanner data, several are likely to start using scanner data within several years and others are considering the use of scanner data. When member states start with using scanner data, they commonly start with that of supermarkets.

This practical guide addresses the use of scanner data for the supermarket assortment of food, beverages and personal and home care products, in short: supermarket-type scanner data of fast moving consumer goods. This means that the guide would also be useful for drugstores and other retailers selling such types of goods. It does not address processing scanner data for goods like clothing and electronics where the assortment changes very frequently. In the case of electronics more explicit quality adjustment methods may be needed.

Besides the increasing use of scanner data supermarket-type fast-moving consumer goods, NSIs are also looking into using scanner data from other types of retailers – clothing for example – which require methods that can deal with frequent changes of products. The methods described in these guidelines should not be used for such volatile assortments as clothing. The methods described in these guidelines assume a certain stability of the supermarket and drugstore assortment and GTINs or other item codes.

Why a practical guide?

The first reason is that it is **Eurostat's responsibility** to ensure the comparability of the HICPs of Member States. The introduction of scanner data risks introducing incomparability if NSIs develop their own methods for processing the data. As more NSIs want to start using scanner data, it is reasonable to learn from others and to align with current best practices. This will decrease the risk of incomparability.

Secondly, Eurostat supports the **modernisation of price statistics** with the aim of ensuring 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 answer to these challenges. This guide is one form of support, besides the regular organisation of meetings and workshops and the financial support given.

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

Why now?

Currently there are a few NSIs that have been using supermarket scanner data for many years and a method has emerged that provides reliable indices for supermarkets[1]. As more NSI want to start using scanner data, it seems the appropriate time to make a practical guide.

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 prepared that should also address scanner data. This guide aims at providing practical background information to that drafting process.

This guide describes the situation of 2016 and will have to be updated in the future as the use of scanner data develops and broadens.

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

[1] Currently new methods are being explored to process scanner data that is more 'volatile', i.e. where the assortment changes very frequently as is the case for clothing. These methods and the discussion has not reached the point yet where the results could be incorporated in this document. This may happen in the future.

[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, GTIN (Global Trade Item Number, formerly EAN code). Sales terminals record each transaction and scanner data, as currently used by NSIs, 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 the calculation of a unit value price for each GTIN. Other codes than GTINs can be used and these will be treated in Chapter 3, but the guide will use the term item code throughout.

The data is not necessarily generated or collected with the purpose of making 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, for primary data sources such as the traditional price collection NSIs determine and are responsible for all steps of the 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 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, clothing stores, and many others. Currently, scanner data predominantly replaces price collection in supermarkets, and in particular for the supermarket assortment of food, beverages and personal and home care products. As stated before, this practical guide will restrict itself to these goods.

Compared to traditional price collection, scanner data offers several advantages: it provides information on the actual expenditures for all item codes sold (by the retailer whose data is used), provides price information on actual transactions over longer periods of time rather than one day in the month, excludes products not actually sold and includes certain types of discounts. Scanner data is also a better source of information for the inclusion of new items and products in the HICP than reliance on price collectors Using scanner data holds the promise of improving the quality of the HICP, but also aims at reducing the administrative burden on retailers and saving costs on price collection.

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

Scanner data also has disadvantages. A greater dependency on the retailers to provide the data is one disadvantage. Another is that there are methodological issues that need to be addressed. The advantages and disadvantages will be discussed further in Chapter 2.4.

The guide treats what data to ask for, from whom, with what frequency and gives guidance on checking the quality of the data and how to process the data. The guide does not address **how** to get the data as this depends on institutional and legal arrangements at the national level, national customs and the negotiating skills of those concerned. Based on the experiences of the NSIs that have implemented scanner data one can state that maintaining good relations with retailers is of paramount importance to ensure a continued delivery of data and timely solution to issues that – inevitably – crop up.

The focus of the guide is on using scanner data from the supermarket assortment of food, beverages and personal and home care product for the production of the monthly HICP.

Using and obtaining scanner data from other sources or for other statistical purposes, like the Food Price Monitoring Tool (FPMT) or Purchasing Power Parities (PPPs) are not treated here.

Outline of the contents

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

Chapter 3 will discuss the main sampling dimensions: regions, outlets and time. This will define the particularities of the data to ask for from retailers. Chapter 4 will then deal with obtaining scanner data.

Chapter 5 will deal with items codes (GTINs and other codes commonly used).

Chapter 6 will give an overview of processing scanner data. Chapters 7, 8 and 9 will then deal with the processing in more detail: mapping, the static and dynamic approach and integration of scanner data in the overall HICP.

2. Scanner data and the HICP

In this chapter we will compare scanner data with price data collected in a traditional manner and then assess scanner data with a view to the fixed basket principle of the HICP. The chapter will close 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 split up using National Accounts data and other sources, like household budget surveys to lower level ECOICOP aggregates, with each aggregate having its own weight. At the lowest level, there are the elementary aggregates (EA) below which no weights are available.

Within EAs product descriptions for items are made for which prices are collected in shops. Product descriptions are relatively wide (e.g. Jam, strawberry, 150 – 300 grams) to ensure that one and the same item description can be used for a longer period of time and across different retailers.

The diagram below summarises the structure.

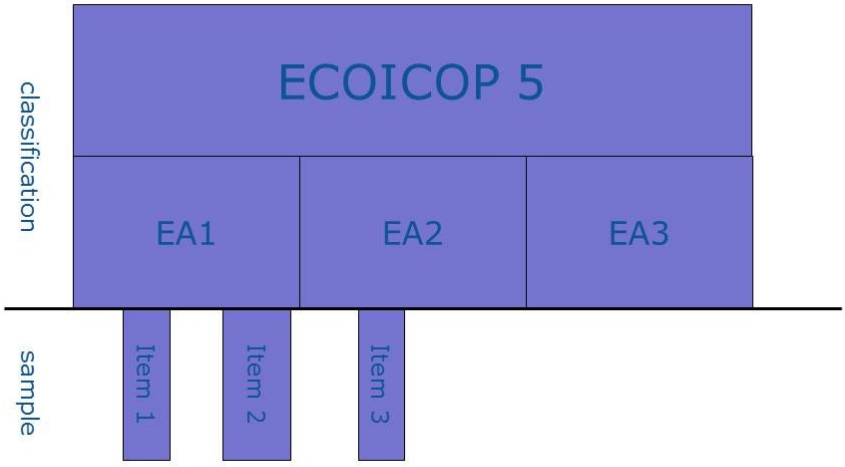


Figure 1 Traditional Structure

The purposive sampling of items and the art of making item descriptions was often based on common sense and experience. In a traditional price collection prices for at most a few hundred products are collected in supermarkets each month for food, beverages and other daily necessities like personal and home care products. With the exception of seasonal products, 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. Traditional price collection draws a more limited sample of prices for the universe of all transactions, and often little information is available below the level of EA on quantities and weights.

Scanner data, for a specific retailer and time period, is an exhaustive listing of all item codes sold, their turnover and the quantities sold. It should be kept in mind that with scanner data the universe of transactions does not change, but the quantity of information available increases dramatically.

Scanner data provides the index compiler in principle 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 recording what was *actually* sold and the inclusion of many more items in the HICP than is feasible with traditional price collection. It also means that individual items could be weighted as turnover information is available. Figure 2 summarizes 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 certain EA need to be used (see Chapter 8), but it is clear exactly what is being left out (the red part).

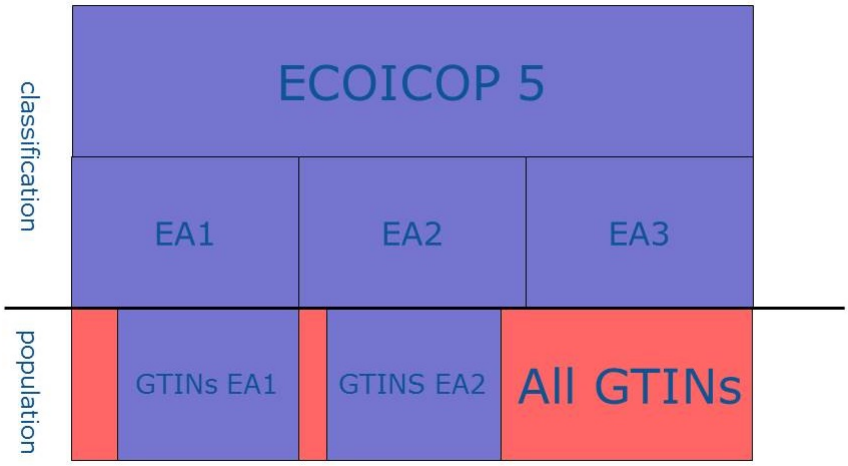


Figure 2 Scanner data

Item codes, e.g. GTIN, identify a good very narrowly so that two goods with the same item code are identical[4] from a consumer perspective. The resulting unit value price per item code is the average of prices ***actually*** paid by purchasers for products, including any taxes less subsidies on the products, and after the deduction of discounts from standard prices or charges. Scanner data does not contain prices for ***product offers*** i.e. the price at which the

[4] This point is not always valid for SKUs, but this does not invalidate the point made here.

consumer is offered the product, which is often an approximation of the price actually paid if discounts are not taken into account[5].

Scanner data reflects the dynamics of actual purchases in each EA 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 recorded and visible in the data set. We will call this 'churn'. The disappearance of items is known as attrition and the rate varies across countries between 25 – 60% of item codes per year. Besides the introduction of genuinely new items, items are often replaced by new versions, called 'relaunches' which are essentially the same item but with some superficial difference like packaging and a new item code. Likewise discounts (20% more) also receive a new item code. In other cases, replacements are more substantial for example when products of a certain brand are replaced by similar products of another brand.

The existence of relaunches, discounts, replacements and genuinely new items suggests that figure 2 is actually more complex. Within the churn of item codes, we group together items with their relaunches and with discounts. When identified 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 attributed to the EA 'Jam' that is part of the ECOICOP 01.1.8.2 (Jams, marmalades and honey).

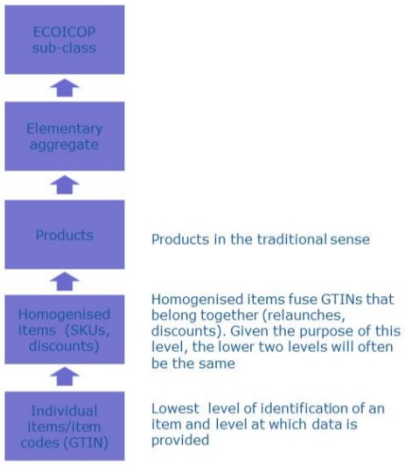


Figure 3 Structure

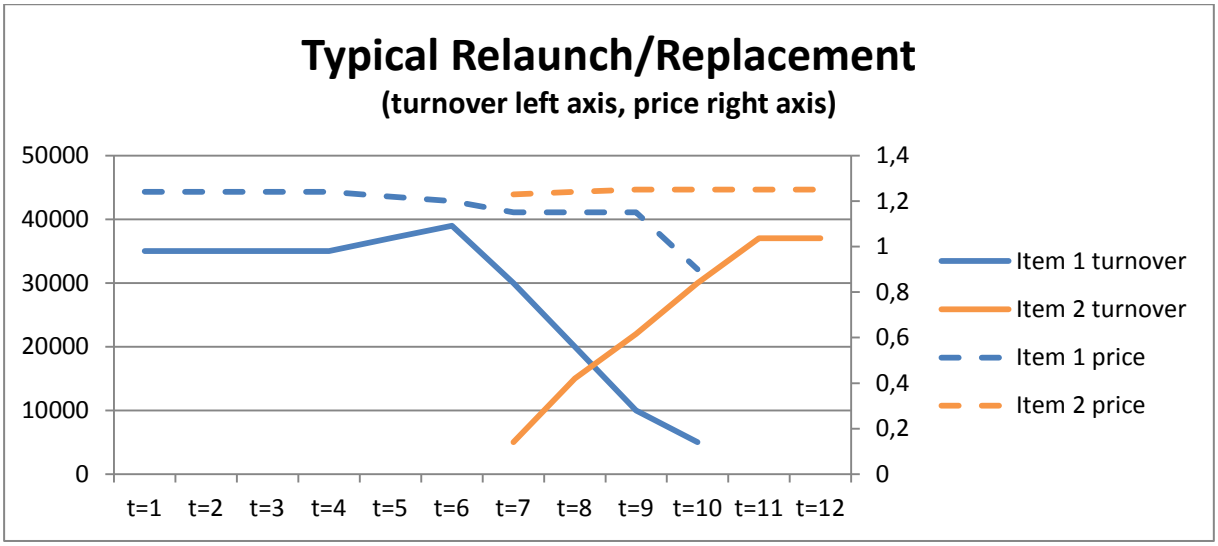
This guideline concerns supermarket scanner data, where the churn is relatively limited compared to clothing. The methods developed in this guideline attribute item codes directly to EAs and not first to a homogenised item. Hence special care has to be taken to capture relaunches and discounts if and when they do occur.

[5] The term 'approximation' is apt because often not all sales in the period that the discount is valid are at the discount price.

2.2. Scanner data and the fixed basket

Maintaining representativity of the sample and dealing with replacements is a challenge. Scanner data offers the opportunity to solve these issues - but not without difficulty.

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 and that of the new item increase.



Graph 1 Replacement

The graph above demonstrates the advantage scanner data offers by providing historical information, and at the same time it offers a challenge: in 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 (fictive) code and a unit value calculated over both item codes. Replacements should be quality corrected if needed, for example if there is small difference in quantity of the contents (186 grs compared to 180 grs).

To maintain representativity, products should not only be present in the new month but also be representative i.e. have sufficient turnover/sold in sufficient quantities. One NSI monitors item codes that sell less than 50% of the number sold in the price reference month December and whenever necessary or possible, replace 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 maintains representativity by including products that fulfil certain conditions (average turnover over a number of preceding months together with a specific ranking in the EA). The first system 'pushes out' less representative items, whereas the second system 'pulls in' more representative items.

In these systems seasonal items are treated differently, taking into account cyclical patterns and non-availability in certain months.

For supermarkets, package size adjustments (PSA) may be needed if the brand and type of product remain the same and direct comparison should be used. In other cases a unit value (see above) or bridged overlap may be used if this gives a correct overall price development.

In the traditional system, without any information below the level of EA (e.g. jam) we assume that there is an item described as: Jam, strawberry, 150 – 300 grams. This is not an unreasonable item to include in the HICP given that supermarkets always sell such jam and strawberry jam is always represented. 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.

The representativity of the sample is maintained for an EA (e.g. jam) when using scanner data by having 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.

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. Potentially, scanner data makes replacements visible, where in traditional price collection these may have remained hidden: the price collectors could switch unnoticed between item codes that fall under the same traditional product definition. Moreover, price collectors only see what is available in supermarkets, not what is representative. However it can be a challenge to design systems for detecting replacements at item code level, especially in automated production systems.

The HICP is a Laspeyres-type index and it is the EA that is fixed, e.g. ‘jam’. The weights of EAs remain unchanged throughout the year. The dynamics involved with continuously changing population of item codes in scanner data takes place below the level of EA. Hence, the use of scanner data does not infringe on the Laspeyres principle underlying the HICP, as a matter of fact scanner data could allow the level of EA 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 item within an EA is not left to the price collector, but can be based on objective criteria such as actual quantities sold and/or turnover. The same point can be made for replacements; rather than being dependent on what price collectors may or may not see, scanner data allows for a more exact tracking of the rate of replacement on the market.

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 EA. This keeps these

methods consistent with traditional methods where prices are not weighed. The use of weights derived from scanner data at the EA or higher levels is treated in Chapter 9.

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 (it could hence also be strictly controlled). With scanner data the dependence on the retailer increases substantially. A common reaction is twofold: both improve the relationship with the retailers and, secondly, make 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 scanner data. Such objections should be evaluated carefully as the regulations were not made with scanner data in mind. The processing of scanner data suggests (see 9.2) extending the period for which missing item codes are imputed to 14 months. This is not compliant with existing legislation. However, traditional item descriptions do not equate with item codes. New item codes (possible relaunches and replacements), are included, automatically in some cases. 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 prepared in the near future will have to take the use of scanner data into account.

The last objections could come from deeply ingrained habits and beliefs. Scanner data replaces a long tradition of sampling, price collection, and ways of validating the collected prices. This requires a different, new mind set and this may not always be easy to change. Given the limited number of prices collected using traditional methods, it is possible to check each and every price. This can lead to a certain attention and care for detail; a control of every price. Scanner data cannot be controlled at the same level of detail.

A cost/benefit analysis should be made and the outcome might differ for each NSI.

3. Sampling dimensions

The starting point when deliberating whether or not to use scanner data depends on the answers to the questions: 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 the next chapter.

The target is a monthly index at some level of aggregation with regions and outlet (types)[6] appropriately weighted. Commonly data is aggregated per week and unless selected outlets are sampled, outlets are aggregated to a regional or a national level. Nevertheless scanner data could also be processed per outlet and per day.

3.1. Regions

Larger countries may (wish to) produce 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. Especially for smaller countries this seems a viable option if retailers use national pricing. If retailers use national pricing, then aggregating scanner data over all regions by the retailer is possible, but the validity of this assumption needs to be checked regularly with the retailer.

If a retailer offers *different* items across regions or uses regional pricing, scanner data should be aggregated per region. The reason is that if data would be aggregated to a national level, items sold in *all* regions will have a higher chance of being included in the calculation than items that are important in only one region. We assume here that only items that pass a certain threshold are included in the calculation. We will return to the use of thresholds later.

To give an example: imagine a country where regional producers for dairy products have a substantial market share in particular regional markets. If data is aggregated to a national level and that data is used for all regions, then such regionally important products may be excluded from the regional price development because they are not important at the national level.

[6] It can be that outlet types (supermarkets, bakers, butchers) are weighed and/or that individual outlets are weighed.

3.2. Outlets

Many retailers operate several formulas or concepts, e.g. Carrefour Hypermarket, Carrefour Market and Carrefour Express. It is not always clear whether scanner data can be aggregated or not over these formulas. The service level offered between formulas, e.g. in terms of opening hours or assortment will often differ and this may be expressed in the price levels. If scanner data is aggregated over outlets these should be homogenous in the service level that they provide. Otherwise, shifts in the population of outlets could bias the index.

Outlet-type replacement/substitution within or between retailers should be treated as they would in a traditional sample: the replacement needs to be found in the same segment or retailer[7]. If scanner data includes all outlets but does not identify which scanner data belongs to which outlet, then changes in outlets (closures, openings, change of formula or retailer) are not visible and these dynamics are included; hence different formulas of a retailer should not be aggregated.

If outlets of the same formula are free to set their own prices then scanner data needs to be supplied at the outlet level.

3.3. Time

Aggregating scanner data over time addresses a new issue; in the past collecting one price per month was often considered enough. Products with volatile prices, like vegetables, could be considered an exception and prices were collected weekly. Scanner data offers the opportunity to process data referring to shorter time periods. Assume scanner data is provided per day. The key question is: is the quality of an item dependent on time? If the quality of the item changes according to the day of the week, then the scanner data should be processed per day. If per week then a week is an appropriate time frame.

If the day of the week or time of the day is a quality aspect of the product and the price set accordingly, then this aspect could be taken into consideration[8]. However if this is a cyclical pattern ('prices always higher in 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 there is no difference here with the treatment of other products where prices can vary depending on the time of consumption either in a fixed pattern (e.g. electricity, telecom) or a more flexible pattern (e.g. flights, package holidays, hotels).

[7] As there are many possible cases no general guidance if given. On a case-by-case basis pragmatic solutions will have to be found.

[8] It should be noted that taking the day of the week into account as a quality aspect is new for supermarkets.

The time period should align with the pricing policy (discounts) of the retailer, so that price changes can be monitored.

The longest interval of time scanner data can refer to is a month. In principle as many days as possible should be included[9], but no days should be included that refer to other months.

Scanner data is usually available for any interval of time: per day, per week, per month. It is important to ensure that the **delivered time interva**l is defined in the same way throughout the year.

Most commonly scanner data is collected per week, i.e. all transactions that take 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. It should be noted that the concept of unit values requires homogeneity of products. It can therefore be an advantage to have well-defined and homogeneous products groups or consumption segments. It can also be necessary to define new product groups with the introduction of scanner data. Currently no specific treatment of different points in time, in line with current HICP practices.

A further consideration is how the delivery of the data fits into the HICP production cycle (including the publication calendar), especially when replacements and quality adjustments have to be made by office staff. Processing data per week 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 all the weekly scanner data files have been delivered before checking for replacements. Collecting monthly scannerdata is unpractical to fit into the production process given the time constraints on the production process.

[9] 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, pages 169-172.

4. Obtaining scanner data

After determining the requirements in terms of regions, outlets and time an NSI can approach retailers with a clear wish-list in an attempt to obtain scanner data. Whether or not the NSIs wishes can be met or a compromise is acceptable is something this guide will not address, save for three remarks:

1. Obtaining scanner data is dependent on the legal and institutional arrangements in each Member State and the relationship the NSI has with retailers.
2. Obtaining the data may be a lengthy process. As a relationship of trust with a retailer grows, changes in data and delivery might well be possible.
3. The wishes of the NSIs 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. It is recommended to obtain scanner data directly from the retailer selling the products to consumers.
   The data should be provided by the retailer as it is data about their economic activity and they are legally obliged to report on these activities. If third parties like regulators or market researchers[10] 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 the correctness of the relevant price indices rests with the NSI and the retailer, not the third party.

2. It is recommended to collect data at item code level. See Chapter 4 for a discussion on item codes.

   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 (e.g. VAT, excise) and further information that identifies the item such as a brand name, a short product description and, if available, the code of a retailer specific classification (RSC). The RSC is a classification owned, maintained and used by the retailer. The more information that can be obtained the better.

   If the retailer uses other codes like SKUs these should be considered for use as primary identifier of an item, especially if they are more stable than a GTIN, i.e. refer to an item under a more general description. Nevertheless it is recommended that the GTINs should still be supplied.

[10] Unless the solution of the NSI is to outsource or sub-contract part of the work or even the actual provision of data.

The data should contain a reference to the period to which it refers, the total turnover and total number of quantities sold. This is important for quality adjustments (changes in contents being quite common for food and beverages). It is also important for the HICP-CT to know to which unit the quantities refer (pieces, kg, litres etc.).

It is essential that information is included that identifies the product. The more information is provided the easier one can identify products and thus process the data either by identifying replacements or grouping homogenous products together. This is the reason why including a retail chain proprietary 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 to their regular counterparts is useful; often a SKU is used for this purpose.

3. It is recommended to collect scanner data aggregated per 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 per week – will allow using 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 of the last week of the previous month. If scanner data can be given per day a more finely grained approach can be taken that would allow the inclusion of days where the week is split over different months.

   Using average prices over three weeks seems sufficient for the HICP in most cases.

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

   Collecting data for every week – but just using the first full three weeks in the month – may be easier for the retail chain to supply that just delivering the first three full weeks (see 5 below).

4. It is recommended to collect scanner data per outlet or aggregated over outlets that are homogenous in terms of the service offered (often a certain concept or retail formula) or per region depending on national circumstances or requirements. Especially for larger countries where products and pricing policy may differ between regions a regional aggregation should be considered.
   Retailers may have different types of outlet (concepts or formulas) e.g. large supermarkets, small supermarkets, mini markets in centres of towns and internet shops that all offer different levels of service. These different levels may have different pricing policies and the numbers of stores of these different types may change as well.
   If pricing policy is determined per region, a regional aggregation may be included, especially if scanner data has to be integrated with traditional price collection (i.e. for

bakers, butchers etc.)[11]. For these reasons, it is recommended to receive data per outlet. However, this data could also be aggregated to geographic regions by the retail chain depending on the concrete situation but the different outlet types or concepts should be kept separate.

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

5. 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. It is recommended to automate the delivery process as this simplifies the delivery of the data to the NSI and thereby also reduces the risk of errors.

6. The details of the provision of scanner data should preferably be laid down in a formal agreement. The HICP is an important statistic. Therefore the production process has to be well organised 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 example contracts[12].

7. It is recommended to use a quality framework suitable for scanner data.
   The use of a quality report for scanner data is strongly recommended as it explicitly and systematically evaluates the quality of the data against requirements. Beyond this primary goal, it could be useful for the following purposes:

   - as a checklist when obtaining scanner data: what topics to discuss with the retail chain or retailer;
   - as list of issues to deal with in a formal arrangement with the retail chain;
   - as (meta) documentation for internal users;
   - as a tool for monitoring 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 is focussed on the output of statistics. Eurostat currently[13] does not have a quality framework or report dedicated to input data, let alone scanner data.

[11] 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.

[12] To follow. It should be kept in mind that the agreements discussed here settle the details of the data delivery, not the legal requirement as such of the retailer to report on their activities.

[13] Work is currently being undertaken to broaden the ESS quality framework and reports to include input data and the statistical processes. Please also see: United Nations Economic Commission for Europe - UNECE

On the next pages an example of a possible quality report is given.

For the general structure we have split the framework into the following three dimensions as proposed by Daas in Secondary Data Collection: source, metadata and data. The dimension source focusses on the supplier and the delivery of the data. The dimension metadata contains indicators on how fit the data is in principle for statistical use. The final dimension concerns the data itself. The three dimensions together cover all quality aspects relevant for 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 the received data with quantified criteria and monitor their development 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.

(2011): Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices. New York and Geneva

Quality report for scanner data

Source

The following first set of quality indicators apply to the source of the data and address some general aspects of the relation with the retailer.

DIMENSIONS	QUALITY INDICATORS	MEASUREMENT METHODS
1. Retail chain	1.1 Contact	- Name of retail chain/data source - Retail chain contact information
2. Relevance	2.1 Usefulness	- Importance of source for NSI (market share e.g.)
3. Security	3.1 Security	- Manner in which the data is securely sent to NSI
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
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

Metadata

The metadata criteria indicate how clearly the coverage is defined and how do 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 does the retailer do?

DIMENSIONS	QUALITY INDICATORS	MEASUREMENT METHODS
1. Clarity	1.1 Delimitation of the retailer	- Is it clear to exactly which parts (outlets, divisions) of the retail chain the data refers.
	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. Data treatment	2.1 Checks (by retailer)	- Population unit checks performed - Variable checks performed - Combinations of variables checked - Extreme value checks performed

Data

The indicators in this dimension concern the data actually delivered. Quality indicators in this dimension should be monitored with each delivery.

DIMENSIONS	QUALITY INDICATORS	MEASUREMENT METHODS
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
2. Accuracy	2.1 Inconsistent objects	- Extent of erroneous objects in source - % of item code codes with illogical relations to (aggregates of) objects. For example: % of item code codes with missing values for variables, % item codes with incorrect VAT rates and % of item codes with illogical prices,
3. Completeness	3.1 Coverage	- Absence of target item code (missing objects) in the source - Check on the number of item code 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
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

5. Item codes

This chapter will present the different types of codes that can be used and will address some general points regarding the sample of codes.

GTIN (Global Trade GTIN)[14] is the current name for the code formerly known as EAN and the most commonly used code when dealing with scanner data. Besides the GTIN, the following codes may also be used: PLU, in-store and retailer specific codes, called stock keeping unit (SKU).

Figure 1 GTIN and PLU

The PLU (price look-up codes) codes are short and used by cashiers or customers when they have to easily enter a code for the item in a cash register or another system that prints a sticker with a code on it that can subsequently be scanned. The PLU code in figure 1 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.

In-store codes are GTINs with a prefix between 20-29 and are only valid for a given retailer. They refer to products that are given a code in the outlet. After weighing an apple as shown in figure 2 (PLU code 3293) an in-store code is printed, the last part of which encodes the price to pay: € 0.81. The last digit is for control. The part of interest to the price statistician is the first part (2306803) that 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 derived. Alternatively the weight or quantity can be encoded in the last part and the price is calculated at the check-out. Hence it is important to understand, for each retailer, what exactly the data refers to.

Figure 2 In-store code

Besides GTINs some retailers may use propriety numbers, which we will name stock keeping units (SKU). These codes can be slightly more generic than GTINs. See figure 3 for an example of a product that does not change for the consumer and therefore has one SKU but two GTINs. A reason could be that the product is produced in two different factories and the

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

producer wishes to distinguish between the factories. Starting from week 39 the same SKU is sold under multiple GTIN codes for a few weeks

Week	RSC	Product description	Unit	Units	Turnover	GTIN
37	1234567890	Chocolate Brand x – 40 pieces	0,375	380	2755	#8000565755675
38	1234567890	Chocolate Brand x – 40 pieces	0,375	561	3540	#8000565755675
39	1234567890	Chocolate Brand x – 40 pieces	0,375	1289	7657	#8000565755675 #8000508890089
40	1234567890	Chocolate Brand x – 40 pieces	0,375	763	4288	#8000565755675 #8000508890089
41	1234567890	Chocolate Brand x – 40 pieces	0,375	1128	6757	#8000565755675 #8000508890089
42	1234567890	Chocolate Brand x – 40 pieces	0,375	912	5591	#8000565755675 #8000508890089
43	1234567890	Chocolate Brand x – 40 pieces	0,375	621	4229	#8000565755675 #8000508890089

Figure 4 SKU and GTIN code

Which codes are used by the NSI as measurement unit will depend on the concrete manner in which the retailer's operations are organised and what data they can and are willing to supply. In all cases the codes should:

1. Identify a unique product. Items with the same item code are identical, but if other codes are used it should be ensured 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 is not useful for measuring price development, because incomparable bunches of flowers would be summed up.

2. Consistently refer to the same product over time. The two PLU codes given above, if they are to be useful, should consistently be used to identify Caribbean Red Papayas and Jazz apples. The lead-time to reusing an item code for an entirely different item is 48 months with the exception for clothes, where it is 30 months. This means that 48 months after the producer sold the last item with a particular item code may be used for another item.

5.1. Business transactions, discounts and returns

The remainder of this guideline will regularly treat the sampling of GTINs. Nevertheless, a few general statements should be made up-front. The current methods of processing supermarket scanner data draw a purposive sample of GTINs purchased by households in a certain outlet during a period falling in the reporting month, after the data set has been cleaned.

The scanner data may contain transactions that should in principle be excluded from the HICP:

1. Scanner data may contain data on transactions between the retailer and other businesses, while 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 that delivers the data. If needed such transactions should be excluded.
2. Scanner data may include purchases of products which are returned within a certain period after the purchase. This needs to be discussed with the retailer, especially how important they are (not for food, but probably for clothing), how they are recorded (in which week) and thus how they influence price development. whether they can be eliminated. This point is also relevant if scanner data is used for weighting.

The importance of the issues raised by both of these points need to be kept in perspective: how important are they really?

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, e.g., discounts of food items which are close to their date of expiration. A related issue that needs to be clarified with the retailer is how coupons are treated in the data files.

These issues need to be discussed with the retailer, but the importance (or lack of) should be kept in mind.

6. Processing Scanner Data: overview

In general terms there are three steps in processing scanner data and two approaches to determine the sample of items. The three steps are (1) mapping GTINs to ECOICOP, (2) sampling a set of GTINs on which to base the EA index and (3) the calculation and integration of the prices or EA indices into the HICP. The two approaches to sampling of items for supermarkets are the static and the dynamic approach.

The mapping of GTINs to ECOICOP is the first[15] step in the processing of scanner data because before a sample can be drawn it has to be clear which GTINs belong to which ECOICOP aggregate. The ECOICOP aggregate will often be below the ECOICOP subclass level (digit-5 level).

After the mapping of GTINs the sampling of GTINs takes place. In the static approach a sample is drawn from the year t and used for 12 months following December. 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 GTINs has taken place, average prices or indices are calculated as needed and then integrated into the HICP.

The ***static approach*** mimics the traditional fixed sample closely: towards the end of a year a purposive sample of GTINs is drawn as representatives for the following year. If an item loses representativity or disappears it is replaced. This all follows traditional methodology, but with the advantage of having full information on actual transaction on which to base choices concerning the initial selection of GTINs and, if needed, their replacement during the year.

This method has advantages in some circumstances. If scanner data has to be combined with traditionally collected data in the sense of adding prices taken from scanner data and adding them to the set of prices collected by the price collectors. In these cases it may be convenient and efficient to 'hand-pick' those GTINs that best fit with the product descriptions used in the traditional price collection. The method also has disadvantages in that it is labour intensive and makes limited use of the available data. In the case of the static approach it may be that prices are added to the prices collected by price collectors, whereas in the case of the dynamic approach indices maybe calculated for EAs.

The ***dynamic method*** eliminates some items as not appropriate (e.g. implausible price changes) and then – automatically – selects all items that pass a certain threshold which ensures that the chosen GTINs represent a specified percentage of turnover. The method ensures a representative sample of GTINs is drawn 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 GTINs that have a

[15] We assume that the completeness and correctness of the data set have been established.

turnover above a certain threshold. The method resembles monthly replenishment and chaining.

The dynamic method is favoured when substantial amounts of scanner data have to be processed. It comes with the advantage of being automated, transparent, and needs no human intervention. However, if relaunches and replacements occur frequently these need to be dealt with separately.

The main difference between the two methods is that the dynamic method is automated to a larger extent which allows it to be scaled up more easily to include more retailers and increase the coverage at no cost. The automatic inclusion of new items should also be considered an advantage.

When processing such large numbers data mistakes, e.g. in the mapping to ECOICOP, are bound to occur and might be seen as acceptable within certain limits. It is hence important that procedures are put in place to check the quality of the automated procedures, monitor the number and nature of mistakes and ensure that the automated systems are improved. Besides the checking of the plausibility of the outcomes – that is already in place in the traditional processing of prices – scanner data requires at least two additional quality checks to be made:

1. Checking the correctness of the mapping of items to ECOICOP. This is valid for both the static and the dynamic approach.
2. Checking whether replacements have been included correctly. This is valid for the dynamic approach in particular.

These quality checks need to be integrated into the production process. A quality report on the processing of the data that focusses on the development and stability of metadata should be integrated into the production process and allow for monitoring all steps in the production, from the acceptance of the data, the mapping, replacements and the plausibility of the outcomes.

7. Mapping and initialisation

The first step in processing scanner data is mapping the items to ECOICOP. The level of product classification to which one maps the GTIN will be the lowest level of classification used by the NSI; often the 6- or 7-digit level constitutes the EA. Mapping GTINs is partly unique to each retailer as the data each of them provides differs; descriptions for the same GTIN need not be the same across retailers.

Given the large number of GTINs, manually mapping each GTIN is not realistic in terms of resources needed and automated procedures are preferred. Often a retailer specific classification (RSC) can be used as a short-cut. It is also possible to simply replace the 6- or 7-digit level with the detailed RSC aggregates. For example, before the introduction of scanner data one NSI defined some 175 EAs in ECOICOP division 01 and with scanner data now defines 400 EAs based on the RSC.

> ***Mapping GTINs to ECOICOP is something new using techniques from semantic technologies, artificial intelligence and machine learning. Eurostat has started a separate project with external partners to develop methodology for mapping. Hence the treatment of the subject is rather summary.***

In the initialisation stage it is important to understand which GTINs should be used and how stable the GTINs are, especially if PLU, in-store and SKU codes are used. The scanner data set may likely also contain groups of products, e.g. clothing, home decoration that are to be excluded or have to be allocated to a different ECOICP division. If GTINs have unclear meanings, then they are not useful and should be excluded or alternatives have to be discussed with the retailer.

It is important to understand the dynamics of the market at EA level. How many items leave and how many new items are introduced? The more frequently this happens the more important it is to ensure that replacements are correctly processed.

The mapping procedure, especially if parts are fully automated, may lead to an incorrect mapping of a few GTINs to ECOICOP. A solution to this is not to build procedures that are 100% water tight because the cost is prohibitive. It is better to monitor the automated systems separately from the production cycle, using a sample of newly mapped GTINs, accept that some mistakes are likely, and use the results to improve the automated mapping procedures. The number of mistakes should, of course, be below a certain threshold. If the threshold is surpassed, corrective action may be called for before using the outcomes for the calculation of 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 GTINs, identifying items and groups of items that are to be excluded and developing automated mapping routines. For each retailer a separate initialisation is required.

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

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

c) The churn of the item codes has to be understood. For this a longer period has to be studied. Do codes refer to the same item each month? What is the attrition rate[16] and the rate with which new item codes are introduced? The results of these tests will determine whether what replacement strategy needs to be implemented, and if so, how.

d) It is recommended to use the retailer specific classification (RSC) and map this to 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 being white rice then one may assume that all items with that classification should be classified in ECOICOP 01.1.1.1 (Rice).

e) It is recommended to identify the items or groups of items to exclude, e.g. item codes with unclear descriptions, e.g. a code could be used to designate the daily changing 'special pastry offer' or 'bunch of flowers'.

Monthly mapping process

f) It is recommended to use the automated mapping process 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 classifying algorithm should be used. If such an algorithm is not feasible, new item codes will have to be classified manually, by visual inspection of the item description and preferably focusing on item codes with a high turnover and item codes sensitive to changing item descriptions e.g. fresh products.

h) It is recommended to perform the following checks each regularly:

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

a. Monitor the changes in the retailer classification; e.g. item codes that move to another group

b. Check the mapping of item codes where the description changed.

c. Check if item codes retain the same RSC/ECOICOP classification.

d. Check item codes that were excluded, for example products with unclear description.

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

i) Ensure that results can be reproduced: store previous versions of the RSC and the mapping to ECOICOP.
j) The quality of the (automated) mapping needs to be controlled. It is, therefore, recommended to randomly select a sample of item codes and check the correctness of the mapping. Errors should then lead to improvements in the (automated) mapping procedures. There are different ways to the checking, e.g. visual scan.

7.2.Mapping with RSC

After a thorough analysis of the RSC, the item codes of RSC aggregates 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 RSC aggregates that could not be mapped, like, e.g., 'Thai food' that could contain rice, vegetables, condiments etc. If a reallocation of item codes would make a significant impact in terms of turnover of the EA where the item code belongs to, a reallocation of the item code should be made. If an RSC aggregate is considered insignificant in terms of turnover it may be left out altogether. Nevertheless it should be monitored, so that if it passes a certain threshold the items are again included.

The mapping should be repeated *each month* for all new item codes. During the monthly mapping, potential replacements could be filtered out by monitoring new and disappearing (decreasing turnover and quantities) item codes.

A further number of checks should be programmed that monitor for changes in the RSC, item codes that change RSC, item codes where the description changes.

If the RSC changes, re-mapping of the RSC to ECOICOP may be necessary. It is generally not in the interest of retailers to change the classification too often, however they should be made aware of the need to inform the NSI in a timely manner if changes in the RSC are foreseen.

7.3. Mapping without RSC

If no RSC is available or it is not useable, alternative methods have to be used. A manual classification of item codes is not recommended due to the resource implications. A better option is the use of classifying algorithms.

Imagine a retailer has ***one*** RSC aggregate for tea, coffee and cocoa which has to be subdivided into ***three*** groups: coffee, tea and cocoa. Considering the aim is to classify text into categories, two approaches to machine learning are available: supervised and unsupervised learning algorithms. The first option is recommended.

With supervised machine learning a pre-classified training dataset is provided, this allows the chosen algorithm to categorize unlabelled test data using the similarities between the training dataset and the unlabelled text. If the test dataset is labelled correctly or if the margin of error is small the model can be used to classify the remainder of the dataset and new products every week. In the example we would have to manually assign a part of the dataset to the three groups (coffee, tea and cocoa) after which the model is tested and evaluated for a part of the unclassified dataset.

With unsupervised machine learning no pre-classified training set is needed, the algorithm classifies the categories based on the data. This is not recommended because in the case of the HICP the categories (=EA) are predefined.

It is likely that a mix of methods will have to be used to ensure that all item codes are 'accounted' for: either they are excluded, automatically linked or mapped with some other method.

8. Sampling: static and dynamic approach

The static and dynamic approaches have been introduced in Chapter 6 and will be described in more detail in this chapter. After providing some background a step-by-step description is given.

It should be kept in mind that one of the main differences is the automated nature of the sampling process in the dynamic approach. The values for the filters used in the dynamic method could be used in the static approach as well.

Generally, representativity of the sample is understood to mean that it should include those items that represent a large share of turnover. It is quite common that a few item codes are responsible for a large share of turnover and these are the items that should be included.

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. However scanner data does provide better information which would allow targeting representativity by each pair of consecutive months (dynamic representativity) by monitoring turnover data and by replacing item codes that are not (sufficiently) representative anymore.

The selection over a longer preceding period is needed to ensure that the sample is representative for the whole year (January to December) and thus suppress typical December sales.

A balance has to be struck between representativity and the ability to maintain the sample. The maintenance is influenced by several interrelated factors: the size of the sample, churn and the efficiency of the system to suggest replacements that are essentially equivalent to the originally sampled items. The filters used monthly in the dynamic approach are also useful to determine the sample in the static approach.

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

Monthly Cycle in the static approach

In the monthly cycle a comparison is made between the previous period(s) 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 a timely fashion. As a consequence 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) It is recommended to fix the sample of items for year t 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 is recommended to ensure that in the initial fixed sample the items cover a sufficiently high percentage of turnover in each EA and the sample is relatively stable. Depending on the number of items and the distribution of turnover over the items a coverage between 50 and 80% is not uncommon. Because the details differ across countries, retailers and EAs no precise guidance can be given.

c) It is recommended to ensure that the resulting fixed sample size is maintainable in terms of personnel resources needed to maintain it.

Monthly production process

d) The validity of the sample (=the occurrence of items in the observation month) has to be checked every month. It is recommended to do this per week, if the data is delivered per week.

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

f) To ensure representativeness of the sample, it is recommended to replace items where the turnover (or number of sales) drops below a specific percentage or value of December t-1.

   The precise percentage is to be determined pragmatically: a sufficient turnover should be covered and a sufficient 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 not representative anymore. An alternative is a method that includes all items codes that pass a certain threshold in the observation period. This makes the method more dynamic, in that new products may enter the index sooner.

h) If prices are the output of the scanner data production process it is recommended that these are processed further according to standard procedures for validation and checking plausibility of outcomes.

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

8.3. Dynamic approach

The dynamic approach was introduced to increase the quality of the HICP without a large increase of resources associated with the manual labour involved in the static approach. The dynamic approach would 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 EA. It is a matter of empirical fact that often many products 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 each month comparing the current month with 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. Hence relaunches have to be treated outside of 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 initialization 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 products where strong decreases in price and turnover suggest that the product will be taken from the market and loses representativity. The aim is to eliminate and the downward effect on the index of clearance prices. Such a filter is needed (despite pt. 4 below) to ensure that products leaving the market are taken out in a timely fashion because the turnover may, even in the second month, still be large enough to pass the low-sales filter.

3. An outlier filter that removes prices that drop/increase above certain thresholds. This filter ensures that clear (decimal) errors are removed, but also items where the discounts are so severe that prices can drop practically to zero. An example is when saved up coupons would allow the consumer to get a product for free. Such outliers should of course be investigated and causes for the error should be 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 (50 – 80%). The low-sales filter 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 number of products in the EA and $\lambda = 1.25$. The value 1.25 for λ is empirically determined so that the selected item codes represent about 80% of turnover. It should be noted that because this method selects item codes based on turnover that certain product segments – low-value or brandless – items may not be included in the sample because their low price also contributes to a low turnover.

The exact values for the filters cannot be given as they depend on markets and vary between countries.

Relaunches and replacements form a potential problem for this method 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 EA 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 products. 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 products (re-use of an item code by the producer for another product) is minimal.

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

Monthly production process

a) It is recommended to develop filters within the initialization process for scanner data from a new retailer.

b) Filter values may differ among countries and across retail segments.

c) Filters that are typically applied are described on the previous page.

d) It is recommended to treat relaunches and replacements by combining old and new item codes and then calculate unit value indices for the combination. Quality adjustments should be applied as needed; especially for changes in package size.

e) The EA index is calculated on the basis of the matched set of representative item codes that are actually sold in two subsequent 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 EA K. The chain-linked index is then as follows (see HICP Methodological Manual, formula 8.13):

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

where K_1 denotes the set of common products in period 0 and 1, K_2 the set in period 1 and 2 and so on. It should be noted that, due to new and disappearing products, the chain-linked month-on-month index does not reduce to the direct Jevons index. Hence it is important to investigate the degree to which the sets K_1, K_2…K differ and the NSI can assess the risks. If the changes are substantial as may be expected for clothing or cosmetics the method should not be used[17].

[17]For an example where this method may 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, pages 50-71

Superlative indices (Fisher, Törnqvist) may not be used in combination with chain-linking as these formulas may lead to considerable drift in the index[18].

Prices for item codes that are not present in subsequent periods are imputed[19] by the price development of the EA for a period of around 14 months to ensure seasonal items re-enter the index at the correct time allowing for shifts between years due to the weather but also holidays as Easter.

The dynamic approach uses a threshold (low-sales filter) and some item codes are not selected for the calculation of the index, but these may pass the threshold in later months. If an item code that was included in the calculation, fails to do some in some subsequent period, it is imputed for at most the same period of about 14 months.

f) A quality control system that is independent of the production system has to be put in place. It should regularly check whether or not the production system accounts for replacements in a correct manner.

There are risks involved in the use of highly automated systems for the monthly production in which algorithms make the choices and where the plausibility of the final outcomes (price indices at some level of aggregation) is the normal check.

To offset these risks the performance of the algorithms needs to be checked regularly and independently because the algorithms assume a certain stability of the population on which they operate (churn e.g.); assumptions that may not be valid anymore at some time.

[18] see e.g. 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, pages 24-35

[19] There is some tension between this recommendation and existing HICP legislation that needs to be resolved in the light of using scanner data. *(ME: In my view, that it an important statement which may be moved to the main text.)*

9. Integration

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

9.1. HICP CT

For the constant tax rate HICP it is recommended to estimate the prices on the basis of available metadata per item code. If such data is not available, the calculation may be based on an estimate of the average of the tax rates are the most typical tax rate applicable to the EA. If this is the solution that is chosen then the EA should be chosen in such a manner to ensure the item code are homogenous in terms of the tax regime applicable, i.e. item code with and without excises should not be mixed.

The CT indices can be calculated at the level of detailed EA indices and this would be the preferred solution when large quantities of prices are processed.

9.2. Seasonality

Seasonal products are those that are not available throughout the year (strong seasonality) or whose prices and weights fluctuate in some pattern that is synchronized with the seasons or time of year and, typically have periods of no sales. In the case of weak seasonality (see below) there are no periods without sales.

Strong seasonality is dealt with in Regulations and scanner data for such items can be dealt with according to those regulations. However, scanner data offers the possibility to take a more finely grained approach as it contains more detailed and up-to-date data. Scanner data makes seasonal fluctuations of expenditure shares and prices explicit, where traditional data sources may not do so at the same level of detail.

Scanner data also allows monitoring the start and end of the seasons more precisely. To ensure the correct application of the regulation in an automated system where items are selected according to turn over the period of imputation needs to be set in such a way that an item can return within the annual seasonal cycle.

Season products can be included from the first month in which they appear. This would allow the index to capture shifts in seasonal patterns. If the EA are defined somewhat broadly (i.e. apples, citrus fruit, berries etc.) then prices can enter from the first month that they appear. If EA are more narrowly defined (e.g. strawberries) then it can happen that the start of the season (in reality) does not coincide with the start month in the aggregation scheme. In the

first case strict-annual weights could be used and in the second case class-confined weights may be more appropriate.

Commonly an imputation period of 14 months could be set for all items codes that disappear regardless of the EA. This means that seasonal items – e.g. relating to Easter/Christmas – could be included accommodating shifts in seasonal pattern.

9.3. Integrating scanner data in the HICP

The integration of scanner data with traditionally collected prices can be done in several ways.

We can distinguish two cases: when scanner data covers a part of consumer expenditure for a specific ECOICOP aggregate and when scanner data from supermarkets covers all of consumer expenditure for a specific aggregate.

In cases where scanner data only accounts for a part of consumer expenditure, the aggregate weight must come from the sources traditionally used. For example, if bread is sold by supermarkets and bakeries, then the total weight for bread would come from National Accounts. The weight for bread could then be split between bakeries and supermarkets using Retail Statistics or other sources.

a. For supermarkets a further breakdown of weights can be made using scanner data.
b. For the bakeries a traditional price collection takes place.
c. A supermarket for which no scanner data is received can be treated as a baker.

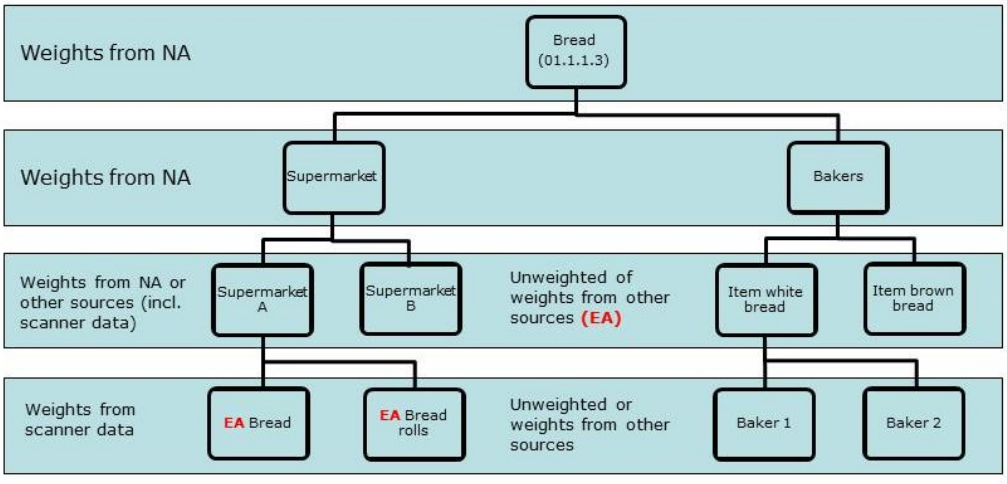


Figure 5 Integration scanner data and traditional data

Note that in this solution all bakery products from a supermarket could potentially enter the index in either of the two EA (bread and bread rolls), whereas for the traditional bakeries the items (white and brown bread) are EAs. The EA for the supermarkets could be different for each supermarket.

If scanner data is not important for a certain EA, but some prices from a scanner data retailer should be added, then the same solution as described above could be implemented, but a few relevant prices could be 'lifted' from scanner data and added to the traditional price file.

The dynamic method should be used in combination with the first method, as the choice for an automated method is not compatible with a traditional way of processing prices.

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