New data sources in the CPI

Large volumes of data is today's price list

After nearly 50 years of price collection via questionnaires, price lists and shop visits, CPI price collection has undergone a significant change. Over the past 10 years, traditional price collection has been replaced by increasingly automated digital data sources, such as web scraping data from web pages and direct collection from enterprises and government agencies via register and transaction data.

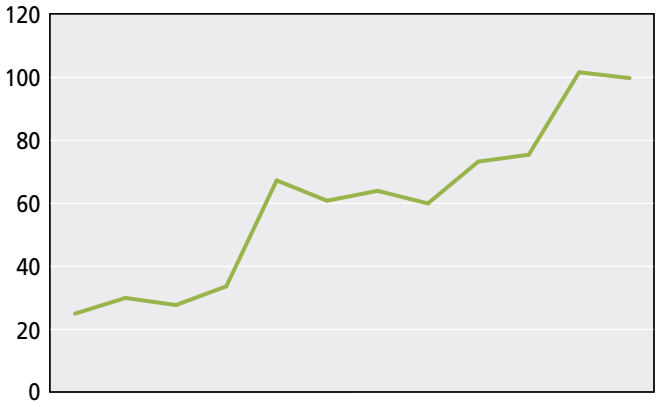
When Stig-Helmer Olsson walks into the Suntrip travel agency, there is already a queue, and the Storch couple from Mjölby are at the counter. In a broad local dialect, Mrs Storch asks the travel agent: "We want to spend our honeymoon on the Canary Islands. How much would that cost?" The travel agent replies obligingly that the price depends on three things: where you travel from, extra charges for peak season and a surcharge for the oil price. The famous Swedish comedy film The Charter Trip (Sällskapsresan) is a vivid study of the purchasing process and pricing of a charter trip in 1980. The purchase was made at a travel agency and the prices were fixed, although a few factors might vary. At Statistics Sweden, prices for package holidays were collected for the upcoming seasons from brochures published twice yearly: the summer and the winter catalogue. In 2020, this market looks very different. Today's Stig-Helmer searches, compares and purchases his package holiday via search engines or on the charter operator's website. The prices normally shift on a daily basis. To capture this daily dynamic pricing, Statistics Sweden uses the enterprises' own transaction data. All prices for every travel combination that is purchased and takes place in the current month are included in the CPI, regardless of the time of purchase. The price of a trip to Crete in Greece in the second week of July can differ greatly depending on whether it was purchased one year ago, two months ago or last minute.

Price collection for the CPI increasingly automated

In today's CPI, the bulk of the prices is collected via automated digital data sources. Large volumes of data are delivered weekly directly from enterprises and from government agencies' data storage, or via web scraping of internet pages. Data collection is automated and the collected information is often comprehensive, which means that Statistics Sweden has been able to increase the samples for the CPI and, at the same time, has been able to reduce some of the response burden. Since 2009, when most of the price collection was done via shop visits, the sample size has quadrupled.

Number of price items per month in the CPI basket

Thousands



Source: Consumer Price Index *Data up to and including 2020*

As the diagram shows, there have been two major changes in CPI price collection, which has enabled Statistics Sweden to expand the sample in the CPI basket. In 2013, the bulk of shop price collection in the food retail market was replaced by weekly deliveries of comprehensive register data from the major supermarket chains. The second increase occurred in 2017–2019, when price collection via questionnaires for services such as real estate commissions, dentist fees, air travel package holidays, train tickets and international air travel was replaced by web scraping and register data.

However, the definition of a price item in the CPI in 2020 need to be qualified. The 100 000 price items registered every month in the CPI in fact represent millions of prices. One price item for a carton of milk in 2020 represents thousands of different transactions that occurred over the month, while a price item of a carton of milk in 2009 represented the listed shelf price at a specific point in time. As a result of individual membership bonuses and volume discounts, customers may pay different prices for the same type of good or service.

Transaction data an important development for price collection

Today, an increasingly large part of price collection for the CPI comes for electronic transaction data. Transaction data refers to comprehensive register data with information on the number of sold products and turnover, by bar code (for goods) or by detailed service content. Transaction data for goods is described as scanner data. The information is delivered weekly directly from the enterprises' data storage. Sweden and the Netherlands have the highest proportion of transaction data in their CPIs among countries in Europe. There are major differences within Europe today, and in

many countries the proportion, if any at all, of such data sources is very small. In several ways, basing the statistics on such detailed and comprehensive data has entailed a revolution in price statistics, above all since it is now possible to use information on actual sales of a product per day, week, or month. For instance, such quantity information at the microlevel means that information is available on precisely how many toilet paper rolls are sold at which price, and whether or not precisely that package holiday to Crete was purchased at a last-minute price. In 2020, transaction data is used to calculate price indices for products such as foods and other consumer non-durables, alcohol, real estate services, train travel, package holidays, dental care and medicine. Statistics Sweden is currently examining whether transaction data can be used as a source to also measure prices on home electronics and clothes.

Using quantities to calculate a price involves a new paradigm in price statistics

In the early 2000s, food prices were collected three times a month via shop visits, and the listed shelf price was noted. The price of a specific type of apple in a specific shop was noted three times per month. As information on the number of sold apples was not available, Statistics Sweden calculated a geometric average value of all the observations. In a geometric average value, the price elasticity is normally uniform, that is, if the price is halved, demand (consumption) increases proportionately. When Statistics Sweden uses transaction data, such assumptions are no longer needed. Instead, the price of apples is calculated with a simple weighted average value of actual prices and quantities over the month.

Example: Different ways of calculating the average price of apples

Week	Price/kg	Amount sold (kg)
1	SEK 10	100
2	SEK 5	300
3	SEK 10	100

Unweighted geometric average value: **SEK 7.94**

Weighted arithmetic average value: **SEK 7.0**

In this example, when the number of sold apples tripled when the price was halved, the geometric average value (the old method of calculation) overestimated the average price.

Web scraped prices online

Statistics Sweden has developed an application that collects real time information from web pages, a method known as web scraping. Web scraping is already used, in part, in the current CPI, and as from 2021, this use will be expanded. To produce price statistics from web scraped data, more information is needed than the price itself. This is why information is also web scraped about the item number, name and characteristics of the product. This enables identification of goods over time and matching of goods of similar quality.

As with transaction data, using web scraping makes it possible to collect large amounts of data in an automated manner. However, information is lacking on the extent to which products on the website have actually been sold. With websites as the only source, it is not possible to ensure that a representative selection of products is followed over time. As a collection method, web scraping may therefore be used in combination with transaction data to ensure that people actually consume what is price measured in the CPI.

The risks linked to large samples

As price collection is growing increasingly automated, some low-hanging fruit would be to increase the sample size at a minor cost, or even carry out a total survey of price development in specific parts of the market. Sample uncertainty would thereby diminish or even disappear. However, decreased sample uncertainty should be balanced against any risks that the price statistics could be biased.

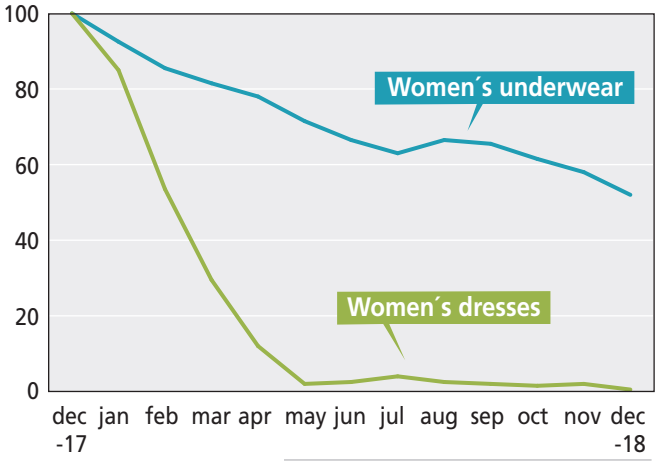
A basic cornerstone of price statistics is that price indices must be comparable over time. If it is not possible to compare products in today's CPI with products measured in the CPI a year ago, the statistics become impossible to interpret. In the CPI, price comparisons must be comparable over time. In a world where the supply of goods changes quickly – new fashion in clothes and new technology in televisions, the basket needs to be constantly updated in order to remain representative of actual household consumption. To maintain the basket's comparability as new products are being added, new products must be matched with old ones of similar quality in the basket. A risk associated with a sample that is too large, is that the amount of matches that need to be handled each month is also much larger. Therefore, large samples of products require the presence of efficient and automated processes for matching products, or alternatively, major resources to manually assess the quality of products in the basket.

The basket shrinks unless new goods are added

What happens if you only price measure products that are constant over time? The goal of comparability may be met, but for most goods and services, this is still precarious. In markets characterised by new fashions or new technology, such as the clothing or home electronics markets, few products are long-lived, and there is a risk that the CPI basket would shrink unless new goods were constantly added. There is a danger that a shrinking basket would distort the statistics, from the perspective of both comparability and representativity, since the basket would have neither the same content over time nor would it represent what is actually being purchased.

The proportion of products in the basket that were sold in the current month, if no new products were added

Percent



Source: Consumer Price Index

Data up to and including December 2018

Different product segments can have varying durations within the same market. The diagram above shows results of an ongoing study on scanner data for clothes[1], which has shown that women's clothes and women's underwear have different durations on the market. After only five months, virtually the entire range of dresses has been replaced. Among women's underwear, on the other hand, there are products that remain longer on the market; about half of the range is still on the market a year later.

When new goods are to be matched into the CPI basket, it is also important to identify the implicit price changes that may occur when products are introduced. It might be a matter of a bag of coffee that has been downsized from 500 grams to 450 grams, while the price per kilo has increased. Around two percent of the foods change packaging sizes over the course of a year. Implicit price changes, or shrinkflation, as it is sometimes referred to, is responsible for around 10 percent of price increases that occur for packaged foods. Although it is difficult to detect the implicit price increases during a shop visit, it is even more difficult to identify them in transaction data.

New data sources give rise to new methods in price statistics

In Sweden's CPI, only a fraction of the large amounts of data collected monthly is used. The reason for this is to maintain control over measurements in order to ensure comparability over time. Furthermore, the monthly information on the number of sold quantities is not fully used in Sweden's CPI.

In recent years, extensive research has been done on various alternative index methods that allow more complete use of transaction data. The methods are pragmatically adapted to handle large amounts of data efficiently, but lack certain theoretical aspects that are met by the traditional methods in price statistics. So far, there is no international consensus on which alternative methods are the best. Statistics Sweden is involved in an initiative led by Eurostat to draft practical guidelines on how to use these alternative index methods (also known as multilateral index methods) in the production.

Price statistics evolve and new data sources give rise to new methods. Old approaches and methods are replaced and statistics are becoming increasingly accurate. At the same time, society is also constantly evolving. When Stig-Helmer undertook to purchase a charter trip in 1980, the prices were fairly stable. Measuring the price development by noting a fixed brochure price gave a correct estimate of the price development at the time. Today, the prices are increasingly dynamic, and a prerequisite for being able to measure the price development is having access to more sophisticated data and relevant methods.

Contact person: John Johansson, +46 10 479 40 12

1 Bubuioc, R. and Tongur, C. (2019) "Preliminary findings in scanner data on clothing". Memorandum to the Council for the CPI (Nämnden för KPI), Statistics Sweden

International comparisons

Comparability of inflation statistics between countries needs to improve

Lately, users of economic statistics have pointed out major differences in price progression between Sweden and other European countries for a number of goods in CPI and HICP. Differences that are due to varying conditions or different choices of method rather than actual price trends make the statistics less useful for international comparisons. Statistics Sweden finds that the calculations of inflation measures could be better harmonised and is pursuing this issue in the European cooperation.

Statistics Sweden works continually on improving the Consumer Price Index, CPI, based on government guidelines and, as far as possible, based on views expressed by the CPI Board and users of the statistics[1,2]. In the past decade, focus has chiefly been national and on issues such as broadened use of electronic transaction data (see the article "Large data volumes are the new price catalogue" in this issue of Sweden's Economy) and improved price measurement of the cost of living in houses and tenant-owned apartments. One issue that has been lurking in the background and which has moved higher up the agenda in the past year is harmonisation between countries.

Several users of price statistics have pointed out that there are considerable discrepancies in price progression between countries within certain areas – a factor that is difficult to explain in any way other than by choices of method[3]. The comparisons in such cases often pertain to HICP (Harmonised Index for Consumer Prices) – an inflation measure that most European countries calculate based on a common framework. The HICP regulations have been developed since the mid-1990s with the aim of harmonising statistics within the EU, and today they only permit a limited number of methods which, theoretically, ought to give relatively consistent outcomes. In Statistics Sweden's view, a gradual harmonisation has occurred continually over a long period of time. If longer time series are analysed, "historical" shortcomings in harmonisation are also evident. Today, the calculation method is much more harmonised than it was 10 or 20 years ago, albeit still with shortcomings. Choice of time period is therefore not insignificant when commenting on harmonisation today.

It is however not certain that a higher degree of harmonisation will be achieved ahead, because access to new data sources, of better quality, varies considerably in Europe. Sweden is therefore currently working more actively for further harmonisation in Europe, primarily within certain specific product areas.

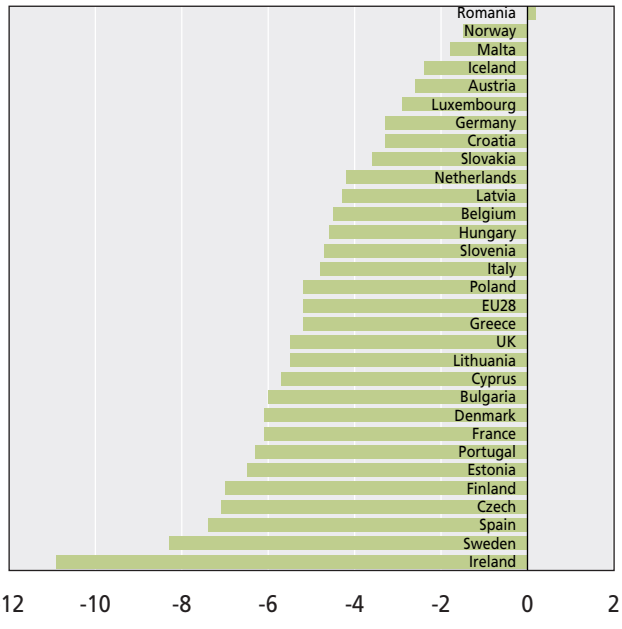
Broad spread in Europe in a handful of areas

The analyses performed both by Statistics Sweden and various users indicate that more distinct shortcomings in harmonisation are limited to relatively few commodities. This refers mainly to TV sets, computers, phones, apparel and pharmaceuticals. To a great extent, these goods are imported from non-EU countries, which ought to indicate a more similar price progression within the EU. Together, they equal approximately 8 percent of domestic Swedish consumption, i.e. of the weight of the CPI basket in 2020.

The chart below shows, as an example, differences in inflation figures, i.e. the 12-month change, between countries for the COICOP group that includes TV sets and computers4. The group has both a relatively large significance in the CPI basket and a large relative difference in inflation figures in Europe.

Average inflation figures, 2009–2019

COICOP 09.1 Audiovisual and photogaphic equipment and data processing equipment, percent



Source: Eurostat

1 CPI guidelines are found in the inquiry SOU 1999:124 Consumer Price Index, and in the Government's response thereto; PROP. 2001/02:1 ANNEX 4, "New guidelines for CPI".

2 Read more on the Consumer Price Index Board here: https://www.scb. se/om-scb/scbs-verksamhet/rad-och-namnder/namnden-for-konsument- prisindex/

3 See e.g. the Riksbank's "Monetary policy report February 2020". https://www.riksbank.se/en-gb/monetary-policy/monetary-policy-report/2020/monetary-policy-report-february-2020/pdf

4 COICOP is a common European classification system and stands for Classification of individual consumption by purpose

Sweden is among the countries that have had the very lowest price progression for COICOP 09.1. It is remarkable that neighbouring country Norway is among the countries that report the least negative price progression. Note that the inflation figures in the chart are averages for a 10-year period and they of course vary between different years.

Difficult to distinguish price from quality

There is great international consensus on the view that a consumer price index should measure the "clean" price progression over time and hence not be affected by differences in quality between new and outgoing products in "the basket". The current guidelines are however less precise on how this should be achieved in all different situations and for each product area.

Statistics Sweden currently uses a handful of methods to judge what constitutes a change in price, and a change quality. All of them are recommended in international manuals, approved under European regulations and have been decided in consultation with the CPI Board in Sweden.

Example – what exactly is the "clean" price change?

In June 2020, Statistics Sweden notes the price of SEK 10,000 for a TV set of Model A that has been on the market for ten months. In July and August, the price of SEK 5,000 is noted for Model A, which is put on sale. In August, at the same time there is also a newly introduced potential replacement, Model B, in the store which is sold for SEK 10,000. In this case, Statistics Sweden needs to determine which of the models are to be included in the inflation calculations and evaluate any difference in quality between the models if a switch is made. Some perceivable options are:

- The entire price difference (i.e. from SEK 5,000 to SEK 10,000) is explained by improved quality ➞ no change in CPI
- Part of the price difference is explained by improved quality ➞ some increase in CPI
- No difference in quality ➞ the entire price difference comes out as an increase in CPI
- The quality is considered to be poorer ➞ more than the entire price difference comes out as an increase in CPI
- No switch is made ➞ if nobody buys a new product, the reduced (sale) price of the existing model is still used in August

In practice however there are a number of difficulties in calculating the "clean" price change for most commodities. Services are often simpler because the fundamental content changes less over time. For example, TV sets undergo a relatively high rate of technological development, with the introduction of new models on the market more or less each month. The price that a consumer has actually paid for a new TV model is relatively easy to observe. It is more difficult to judge whether elements of the generally higher price are due to the new model really being better than the old one, and in that case exactly what is better and by how much, expressed in kronor. In connection with this, features that are for instance "useful" need to be distinguished from those that are model-related in new products. Often, there is also an absence of reliable information on the extent to which newly introduced models are purchased, while at the same time the price is often reduced for older established models when new ones are introduced.

National conditions complicate international harmonisation

For the same group of commodities, such as TV sets for example, several different methods are used in Europe today to assess what constitutes price progression and improved quality. Also, a number of methodological choices are made in practice that can affect the recorded price progression even if the same method "on paper" is used in two countries.

Some fundamental differences between countries are varying resources for producing statistics and legal disparities, which entail different possibilities to collect the data needed to attain good quality in the inflation statistics. Price measurements for the same type of goods are designed in different ways depending on national conditions. It can be a matter of different sample sizes, different rules on when and how product switches are made and of quality adjustments being made in a more or less standardised way. If a country has access to price and sales data for all TV sets from a company, the price progression can be calculated with greater quality compared with "traditional" price collection whereby statistical agency staff visit physical stores and collect prices for a relatively small sample.

The quality adjustments that occupy a price statistician on a day-to-day basis are such that are made during the year, due to individual products declining in popularity or being sold out, and which are thus in need of replacement. However, achieving broader harmonisation of price indexes in Europe requires looking at more than these methods alone. Ultimately, it is the combination of several choices of method as a whole based on different conditions that affects the inflation statistics.

Because the HICP regulations are relatively general in certain areas, national differences – for instance in terms of chaining frequency – still affect how often new samples are drawn, which in turn governs how many switches need to be made during the year. In Sweden, the product and store sample is renewed on a greater scale at the turn of each year, and in connection with this an assumption is made as a rule that the entire price difference between the baskets of both years is explained by a change in quality (see the article Impact of judged quality improvement on price indexes in this issue of Sweden's Economy). Further differences in choice of methodology between countries can also be found in price collection and classification.

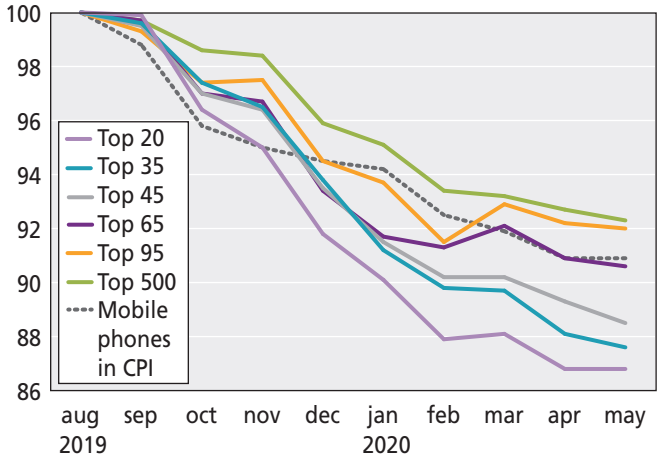
Example 1 – difficult to measure price progression without sales volumes

To illustrate the problems associated with different choices of method, a diagram is provided below with the calculation method monthly chaining for mobile phones. In the Swedish CPI, monthly chaining is used as a calculation method for mobile phones, computers and computer accessories. With the exception of the dotted line, the data in the example is however not based on CPI data, but has been collected during the period August 2019 through May 2020 from a Swedish price comparison site.

In the diagram, different index outcomes are shown depending on the size of the product sample. Despite the relatively short time period (10 months) there are indications that the size of the sample affects the index series. In the diagram below, we see that the index progression is consistently higher with a larger sample. If the 500 most popular models are included, this leads to a higher progression than an equivalent calculation based on the 95 or 20 most popular models. Hence, the larger the sample, the greater the weight of the products that do not sell particularly well and that have a weaker downward price trend. As a reference series, the official index series from CPI has also been inserted.

Monthly chaining for mobile phones: product sample

Index August 2019=100



Source: Consumer Price Index

Data up to and including may 2020

The objective of CPI is to measure prices of products that are actually purchased and, in this case, there is a risk that a large sample would exaggerate the estimated price progression for the target population. In the absence of information on both price and sales volumes, it is in practice difficult to measure the price progression for a representative sample.

Example 2 – prices in one or all stores?

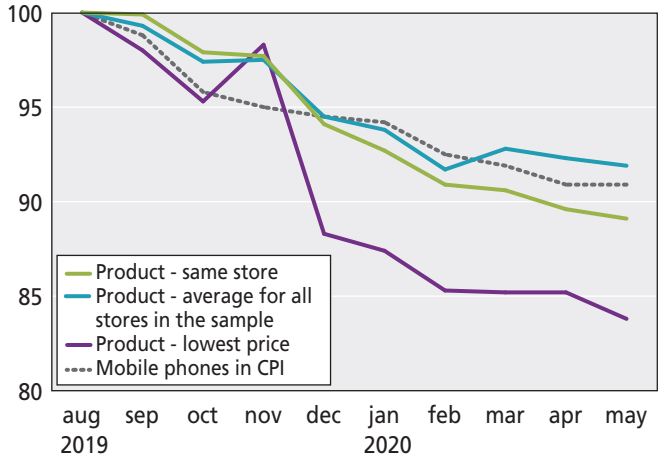
Another methodological choice is whether or not price progression for goods and services should be followed in the same stores over time. It could be argued that an online consumer today has relatively good possibilities of making informed decisions and substituting between stores. This would favour following the average price of a product in all stores in the sample, with varying store weights from one month to the next. Another option, in the absence of weights, is to measure the price in the store(s) that happen to have the lowest price each month. If consumers substitute towards stores with lower prices, this comes out as a price reduction.

On the other hand, it could be argued that there are considerable differences in service level between different stores, and that price effects arising from consumers substituting between stores ought therefore not to be included.

The diagram below shows how the product definition can also affect which price progression is measured. When the measurement of a product follows an average price for all stores ("product – average for all stores in the sample"), the index will be higher than when the measurement follows a specific product in the same store over time ("product – same store"). Today, measurement in CPI is performed according to the latter method. If it instead resembles a rational consumer who, all the time, chooses the lowest price on the comparison site ("product – lowest price"), the price progression will, as expected, be lower.

Monthly chaining for mobile phones: product definition

Index Augusti 2019=100



Source: Consumer Price Index

Data up to and including may 2020

The above are examples of national conditions; differences in practical approach and various fundamental considerations could cause different results in the official statistics.

The documentation available today on different choices of method in the European countries unfortunately does not provide a sufficiently detailed picture of potential methodological differences in the statistics.

Statistics Sweden is pursuing the issue internationally

Within the HICP cooperation, harmonisation is an issue at largely all working meetings. In terms of quality adjustments, in the past few years discussions have revolved around, for instance, cars, for which price progression in Eastern and Western Europe has diverged

During 2019–2020 in the various HICP forums, Statistics Sweden has worked actively to improve the evident problems present in harmonisation between countries.

Issues that Statistics Sweden has pursued within the HICP cooperation

- Sweden has proposed that harmonisation efforts should focus on product groups with a high weight and with the greatest divergence in price progression.
- An in-depth process to identify methods will now be implemented for problematic product groups. Sweden is involved, devising a survey that will be sent to all Member States.
- Sweden has also stressed the importance of following up on whether countries actually implement Eurostat's recommendations. To attain harmonisation in the long run, it is crucial to understand why countries choose not to implement Eurostat's recommendations.

Harmonisation will probably be a standing item on the international agenda in future too. In some countries, electronic transaction data constitutes an increasingly important data source, while many other countries primarily continue to rely on visiting stores. In the Swedish CPI, transaction data currently makes up approximately 35 percent. New data sources often enable observing many more prices and also quantities. This enables the use of better index calculation methods which, in turn, could reduce risks of systematic errors. The countries of Europe have varying economic, legal and cultural conditions for improving statistics and will need to develop at different rates. At the same time, it is important not to forget international comparability – an aspect that needs to remain the responsibility of both Member States and the European cooperation.

Contact persons: Emanuel Carlsson, +46 10 479 48 11 and Peter Nilsson, +46 10 479 42 21