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 Over the past 10 years, traditional price collection has been replaced by increasingly automated digital data sources, such as web scraping data from web pages and direct collection from enterprises and government agencies via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.

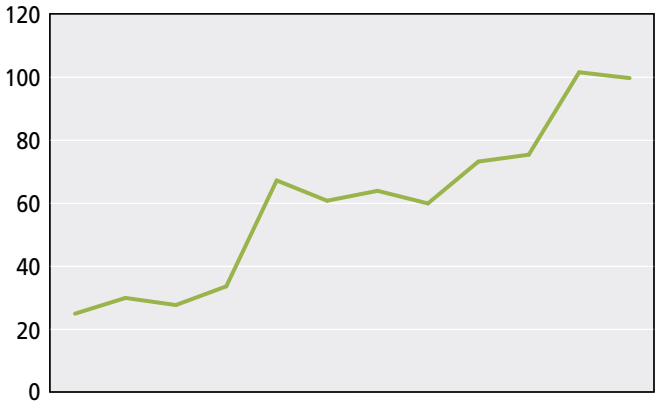
When Stig-Helmer Olsson walks into the Suntrip travel agency, there is already a queue, and the Storch couple from Mjölby are at the counter. In a broad local dialect, Mrs Storch asks the travel agent: "We want to spend our honeymoon on the Canary Islands. How much would that cost?" The travel agent 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.

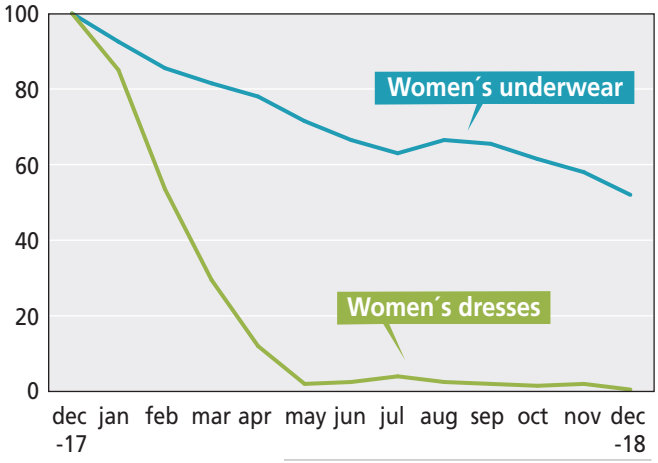
A basic cornerstone of price statistics is that price indices must be comparable over time. If it is not possible to compare products in today's CPI with products measured in the CPI a year ago, the statistics become impossible to interpret. In the CPI, price comparisons must be comparable over time. 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