Consumer price indices: from traditional to new data sources and techniques

Dr Jens Mehrhoff*, Deutsche Bundesbank

EMOS webinar, 18 March 2020

* This presentation represents the author's personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Consumer price indices: from traditional to new data sources and techniques

Presenter: Dr Jens Mehrhoff, Deutsche Bundesbank
E-mail: jens.mehrhoff@bundesbank.de
Date: 18 March 2020

Webinar aims

This webinar aims at introducing price measurement in official statistics, in particular challenging, new areas of measurement. The objective is to introduce participants to the various practical and technical issues encountered while attempting to construct price indices. Questions include:

- What are the conceptual and practical measurement problems in measuring inflation in the digital economy?
- What are the methodological issues and practical experiences with the use of big data for compiling consumer price indices?

Webinar learning outcomes

- To understand the basic principles of consumer price indices and inflation measurement in official statistics.
- To become familiar with the challenges in modern price statistics, including limitations of data and methods.
- To be aware of the need for high-quality data used in monitoring the state and development of the economy.

Webinar content

- The European Harmonised Index of Consumer Prices
- The impact of digitalisation on inflation measurement
- The use of big data sources for consumer price indices

Difficulty level

- Introductory

Prerequisites for the webinar

- A sound understanding of basic statistical theory is necessary.

Further reading and resources

On consumer prices in general

- Eurostat (2018), *Harmonised Index of Consumer Prices (HICP): Methodological Manual*, Luxembourg: Publications Office of the European Union.
- ILO, IMF, OECD, Eurostat, UN, World Bank (2020), *Consumer Price Index Manual: Concepts and Methods*, white cover version.

On digitalisation in particular

- Reinsdorf, M., and Schreyer, P. (2019), "Measuring Consumer Inflation in a Digital Economy," *OECD Statistics Working Papers*, No. 2019/01.
- Quirós-Romero, G., and Reinsdorf, M. (2019), "Digitalisation and the Measurement of Inflation and Growth," 62nd ISI World Statistics Congress, Kuala Lumpur.

On new data sources in particular

- Mehrhoff, J. (2019), "The Value Chain of Scanner and Web-Scraped Data," *Economics and Statistics*, No. 509, pp. 5 – 11.
- Mehrhoff, J. (2018, updated 2019), "Promoting the Use of a Publically Available Scanner Data Set in Price Index Research and for Capacity Building," https://github.com/eurostat/dff.

1. The European Harmonised Index of Consumer Prices

Introduction

- The Harmonised Index of Consumer Prices (HICP) is a specific inflation measure that is developed within the European Union (EU) to result in indices that can be directly compared and aggregated across countries.
- The HICP serves two main purposes:
  - For quantifying the price stability in the European Central Bank's (ECB) monetary policy strategy. The ECB's Governing Council has *defined* price stability as a year-on-year increase in the HICP for the euro area of below 2%. The Governing Council has clarified that, in the pursuit of price stability, it *aims* to maintain inflation rates below, but close to, 2% over the medium term.
  - For assessing the price stability criterion, which is one of the convergence criteria used to evaluate if a country can join the euro area.
- In addition to these specific EU uses, the HICP may be used, like other consumer price indices (CPIs), for economic analysis and for indexing for example contracts and wages.

The HICP concept

- Conceptually, the HICP is designed as a *Laspeyres-type index*. The *short-term* Laspeyres-type index is given by:

$$P^{0t,mt} = \sum_{i=1}^{N} \frac{p_i^{mt}}{p_i^{0t}} \cdot w_i^{0t,t-1}.$$

- The price of the ith product is denoted by p_i ($i = 1, ..., N$), the *price reference period* is December of the previous year and is labelled as month $m = 0$ of current year t, and the *comparison period* is denoted by month m of current year t.
- In other words, each year t is considered as consisting of 13 months ($m = 0, 1, ..., 12$), running from December of year $t - 1$ ($0t$) to December of year t ($12t$).
- *Annual weights*, denoted by w_i, are expenditure shares of the *weight reference period*, i.e. the previous calendar year $t - 1$, and are price-updated to prices of the price reference period $0t$.

Elementary aggregates

- In practice, HICPs are constructed in two stages:
  - a first stage at the lowest level of aggregation, where we have only price information, but not associated expenditure information for weighting purposes; and
  - a second stage of aggregation, where expenditure information for weighting purposes is available for higher levels of aggregation.
- The aggregates that pertain to the first stage of aggregation are called *elementary aggregates*.
- A Laspeyres-type index cannot be calculated at the lower level (stage 1). For stage 1 of the aggregation process, elementary aggregate indices $P_i^{0t,mt}$ are used instead:

$$P^{0t,mt} = \sum_{i=1}^{N} P_i^{0t,mt} \cdot w_i^{0t,t-1},$$

  where $i = 1, ..., N$ denotes elementary aggregates rather than individual products.

Index theory

- Thus, consider an elementary aggregate with a set of K common products in any given time period.
- Many formulas have been proposed in the literature for measuring the price change of an elementary aggregate; the Dutot and Jevons indices are preferred.
- *Dutot index* (ratio of arithmetic mean prices):

$$P_D^{0t,mt} = \frac{\frac{1}{K}\sum_{k=1}^{K} p_k^{mt}}{\frac{1}{K}\sum_{k=1}^{K} p_k^{0t}}.$$

- *Jevons index* (ratio of geometric mean prices, or geometric mean of price relatives):

$$P_J^{0t,mt} = \frac{\left(\prod_{k=1}^{K} p_k^{mt}\right)^{\frac{1}{K}}}{\left(\prod_{k=1}^{K} p_k^{0t}\right)^{\frac{1}{K}}} = \left(\prod_{k=1}^{K} \frac{p_k^{mt}}{p_k^{0t}}\right)^{\frac{1}{K}}.$$

Annual chain-linking

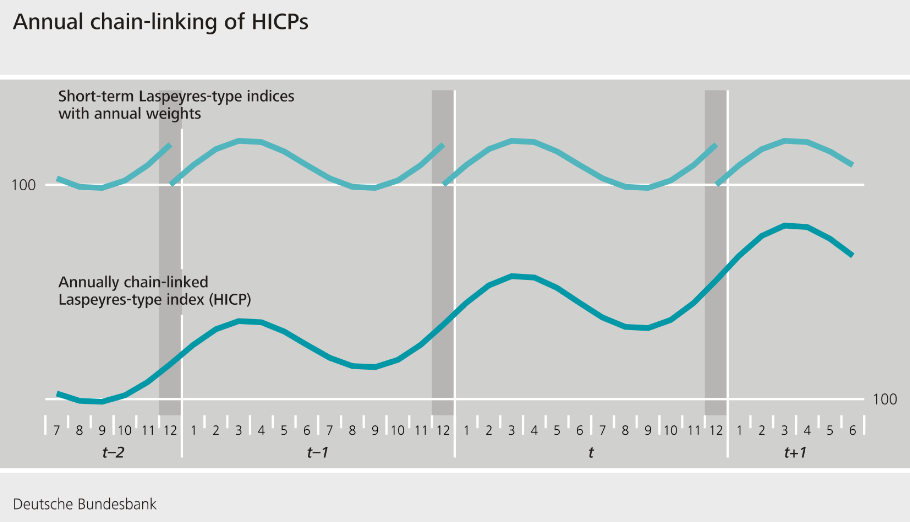
- The Laspeyres-type index compares prices in month m of year t to those in December of the preceding year, $t-1$. When t moves through time, there is for each year a series of 13 index numbers, running from December of year $t-1$ (its index number being equal to 100) to December of year t.
- Now these separate 13-month series can be chain-linked together into a single long-term series, which compares month m of year t to some earlier period. The HICP uses December as the linking month. The *annually chain-linked* Laspeyres-type index:

$$CP^{b,mt} = \Big(P^{b,12(0)} \cdot P^{0(1),12(1)} \cdot \ldots \cdot P^{0(t-2),12(t-2)} \cdot P^{0(t-1),12(t-1)} \Big) \cdot P^{0t,mt} = CP^{b,12(t-1)} \cdot P^{0t,mt}$$

compares month m of year t with a certain year b.

- In this case year b, used in the initial link of the long-term series, is the *index reference period*.

December as the linking month



1. The European Harmonised Index of Consumer Prices

- The HICP concept
- Elementary aggregates
- Annual chain-linking
- (Numerical example – see Appendix)

2. The impact of digitalisation on inflation measurement

Digitalisation and measurement of inflation

- Some longstanding challenges for price statistics of keeping up with a changing economy and capturing quality change have been amplified by digitalisation.
- Whether estimates of CPIs still provide a good measure of inflation in a digitalised economy has become a topic of debate. Several academic and business economists have suggested that digital products are conceptually relevant for understanding consumer inflation as measured by the CPI.
- The claims that prices of household consumption are being mis-measured largely revolve around incomplete adjustment for quality change in products or distribution channels, i.e.,
    1. the treatment of new, and often improved, varieties of existing digital products (e.g. computers);
    2. the treatment of new digital products that replace existing non-digital products (e.g. streaming services replacing CDs); and
    3. improved variety selection of digital and non-digital products (e.g. clothing, books).

Quality change in existing product lines

- Reinsdorf and Schreyer (2019) calculate upper-bound impacts on the deflator (price index) for household consumption in OECD countries.
- In terms of the simulated effects of possible measurement errors based on 2015 weights,
  - the upper bound correction to the index's growth rate for overlooked quality change is -0.41 percentage point, largely driven by the 0.24 percentage point overestimation of the deflator for telecommunication services;
  - the potentially unmeasured savings from digital replacements is -0.11 percentage point; and
  - the upper bound correction for improved variety selection is -0.05 percentage point.
- Combining all the effects, Reinsdorf and Schreyer end up with an upper bound for the potential mismeasurement of digital products of -0.57 percentage point.

New varieties of digital products

- Using price changes of continuing products (matched-model approach) to approximate price changes of new varieties typically leads to an over-estimation of inflation for digital products.
- Sample refreshments help to keep the sample representative and are also occasions for bringing in *new products and product varieties*. In a sample refreshment, a newly selected sample is “linked in”, and the old sample is “linked out”. The first period the new sample comes in is also the last period for which the old sample is used.
- First, in the case of a product undergoing significant quality improvement, the failure to adjust for the higher average quality of the items in the sample being linked in may cause the index to overstate the product’s price change.
- Second, late introduction can lead to price declines early in the life cycle being missed, a problem that is particularly relevant for digital products.

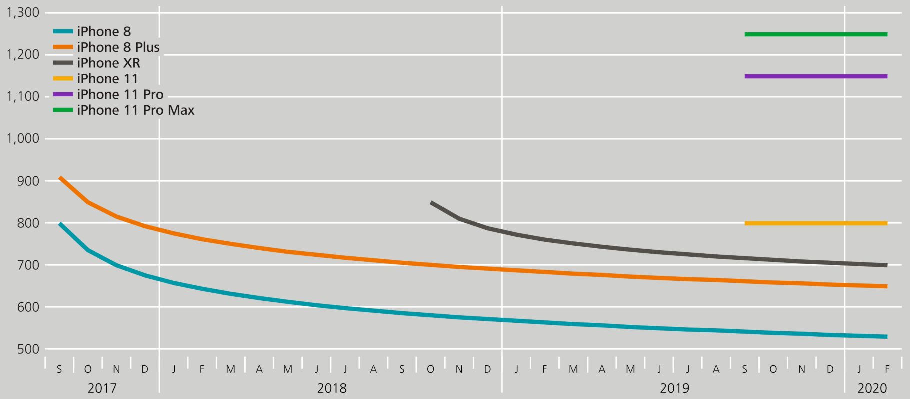
Simultaneous price and quality changes

- However, often quality increases are higher than price differences between old and new varieties.
- Much less investigated, but not to be overlooked, is whether quality change of some digital products may systematically be *overstated*. Examples of quality declines that are not captured in price measures include the requirement to purchase new models of mobile phones and computers in the absence of backwards compatibility of new software with older hardware.
- To summarise, quality adjustment of replacements for existing products within a sample, and of new products coming in during sample refreshments, may miss some quality change. The overlooked quality change and cost of living effects could be in either direction, but for digital products benefitting from new technology, insufficient capture of quality improvements is more likely.

Do higher prices reflect quality growth or inflation?

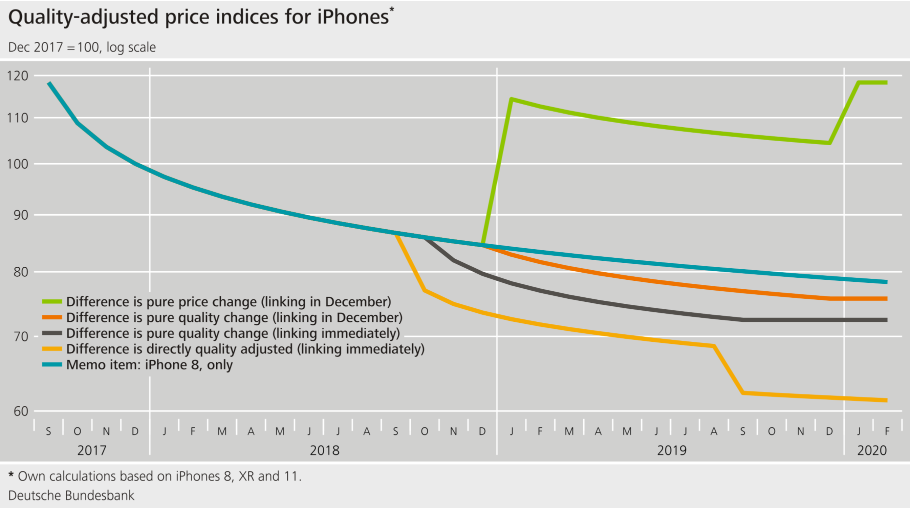
Stylised price paths of different iPhone models*

€

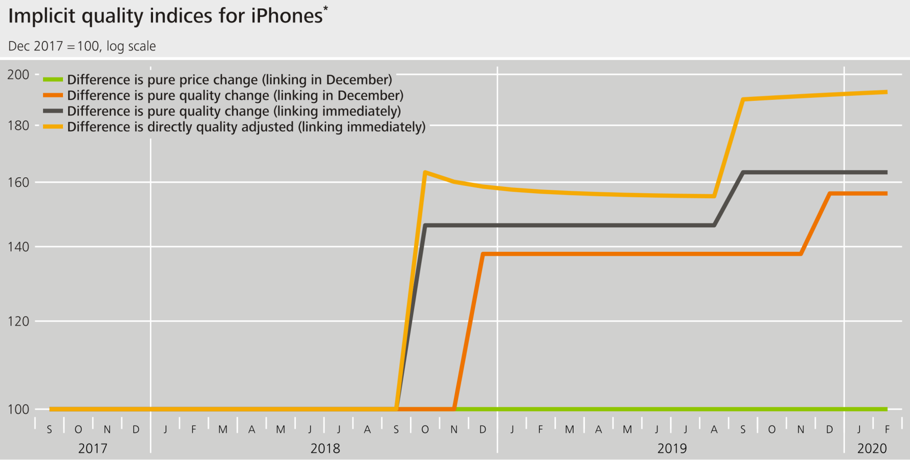


* Own calculations based on list prices by Apple.
Deutsche Bundesbank

Quality adjustment is as much an art as a science...



...and a matter of debate and personal perception



* Own calculations based on iPhones 8, XR and 11.
Deutsche Bundesbank

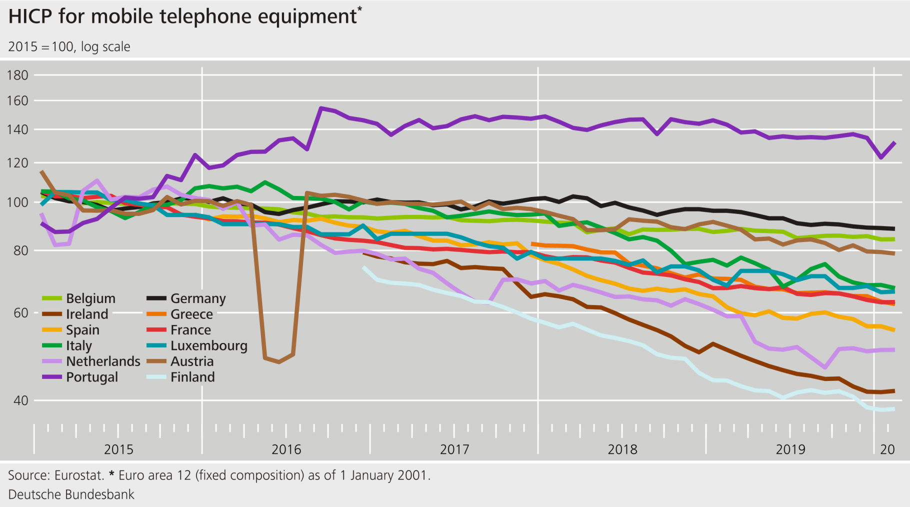
What's the potential impact?

iPhones 8, XR, and 11 compared (from Dec 2017 to Feb 2020)

Price index	Average annual rate (in %)	Quality difference (in %)
Difference is...		
...pure price change (linking in December)	+ 8.07	0
...pure quality change (linking in December)	−12.06	+56
...pure quality change (linking immediately)	−13.83	+63
...directly quality adjusted (linking immediately)	−20.19	+93
Memo item: iPhone 8, only	−10.66	—

NB: Reinsdorf and Schreyer assume a 5-percentage point per year over-estimation of the price change in the affected categories of ICT equipment (incl. telecommunication equipment).

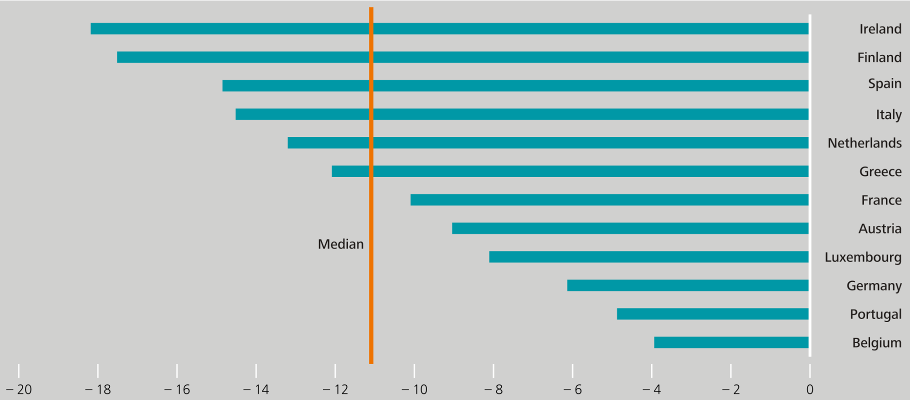
Does the HICP capture reality?



How do countries measure inflation?

Price changes of smart phones*

%



* Average annual rate from December 2017 to February 2020.
Deutsche Bundesbank

2. The impact of digitalisation on inflation measurement

- Digitalisation and measurement of inflation
- New varieties of digital products
- iPhones 8, XR, and 11 compared
- Price changes for smart phones

3. The use of big data sources for consumer price indices

Setting the scene

- With the advent of scanner and web-scraped data, “big data” sources are increasingly finding their way into consumer price indices internationally.
- Scanner and web-scraped data give access to a much broader continuum of products than classical sampling allows. The supposedly better coverage of goods and services comes at a cost, though: churn due to new and disappearing products, i.e. a dynamic product universe.
- Moreover, quantities sold (with scanner data) or at least a popularity ranking (from websites) become available too, thus allowing the calculation of weighted indices rather than the need to rely on unweighted formulae. The cost here is chain drift, i.e. the index might show spurious trends over time.
- The value chain of scanner and web-scraped data can be considering using three stylised phases: i) collecting data; ii) *processing data*; and iii) disseminating results.

Processing data

- There have been several approaches to further break down the second phase, *processing data*. Though by and large similar they differ due to institutional arrangements such as the statistical office's current approach to consumer prices.
- Typical steps include but are not limited to the *automatic classification of products*, intermediate aggregation of "homogeneous" products, rule-based filtering of observations and the *calculation of the final index*.
- The vast amount of products can no longer be *classified* to the official classification of sub-indices for consumer price statistics (ECOICOP) or breakdowns thereof manually but only automatically.
- After the data set has been further edited, the final *index can be calculated*; choices include a fixed basket with a bilateral formula and multilateral approaches in a dynamic product universe.

Classification

- The classification might come from the data owner, at least to some extent. Supermarkets, for example, have their own classification for scanner data which might be useful to this end. The same holds true for web shops, where the products might be presented in a structured way.
- However, should this information not be available or sufficiently detailed for the purpose, one has to rely on supervised machine learning techniques. Yet, this requires the construction of a small labelled data set in order to train the algorithm.
- In addition to information from the data owner, typically product codes (such as GTINs), descriptions (i.e. text) and other metadata (e.g. size) are available. A major challenge in this respect is feature engineering.
- In general, product descriptions are not natural text but use specific vocabularies and rely on different kinds of shorthand. Product codes, on the other hand, follow some kind of a structure.

Example

Example from the Dominick's Finer Foods (DFF) data set

see Mehrhoff (2018), https://github.com/eurostat/dff

DFF category: **bottled juice**

UPC number: **0 14800 00034 4** *Universal Product Code*

0 is the number system digit

14800 is the manufacturer code

00034 is the product code

4 is the check digit (not in the DFF data set)

Product name: **Mott's® 100% Original Apple Juice**

DFF description: **MOTTS REGULAR APPLE**

Product size: **64 oz.** (≈ 1.89 l)

0 14800 00034 4

Index calculation

- There is no consensus on the “right” decisions to be taken if a multilateral approach is chosen.
- Plenty of approaches have been suggested that satisfy circularity, that is the chain-linked index defined as the product of the short-term indices is equal to the direct index, thus ensuring freedom of chain drift.
- The time-product dummy (TPD) method derives the price index from a log-linear regression framework

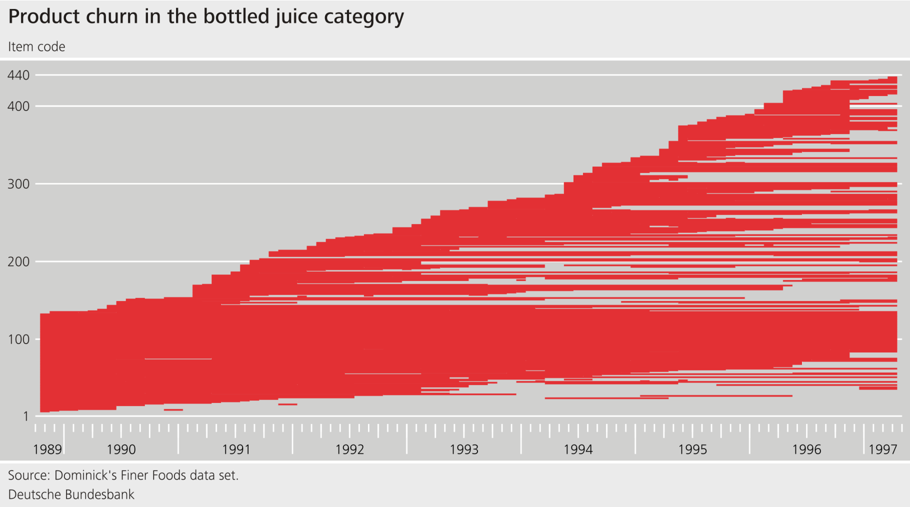
$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t,$$

  estimated by weighted least squares, where the expenditure shares serve as weights.
- The time dummies δ^t are extracted from the output. The price index is derived as the exponentiated time dummy, i.e. $P^t = \exp \delta^t$ and $P^{t=0} = 1$.

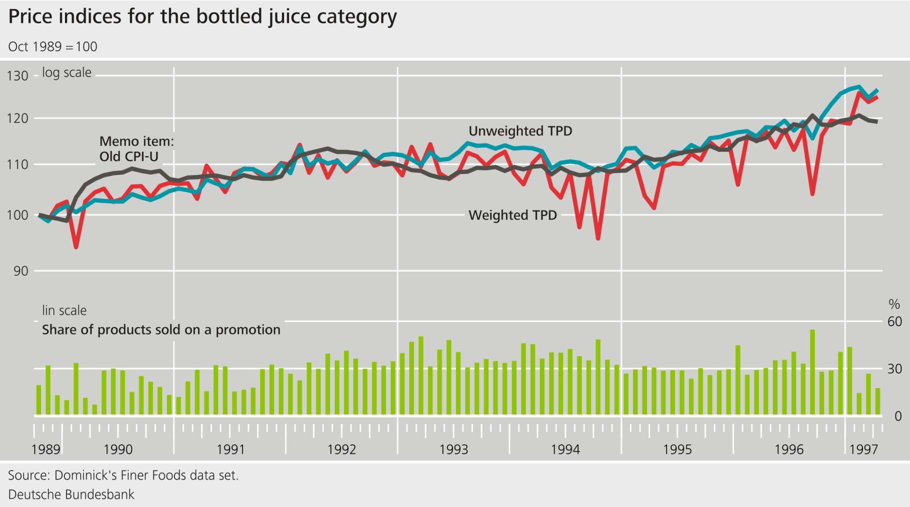
The Dominick's Finer Foods data set

- The Dominick's Finer Foods data set covers 93 stores for 399 weeks from 14 September 1989 to 7 May 1997, and totals 98,884,285 observations of 13,845 products in 29 categories.
- The Dominick's data set is unique for the breadth of its coverage and for the information available on retail prices. Though the data set is historic in nature (and has some gaps), it shows all the specificities one finds in modern electronic transactions data, particularly a dynamic universe due to new and disappearing products (i.e. churn).
- As an example the results for bottled juice are presented, where the weekly store-level UPC ("Universal Product Code") data are aggregated to chain-wide item codes at monthly frequency (from October 1989 to April 1997).
- The within-category average duration of bottled juice products in the sample is 37 months.

Availability of item codes over time



Estimation of price index numbers



Volatility

- Note that the weighted TPD index ends up at 124.9 while the monthly chain-linked Laspeyres index ends up at 1,609.7. So it is clear that weighted bilateral indices are subject to severe drift.
- As it can be seen from a comparison of the weighted and the unweighted TPD index, where the latter is less affected by quantity increases due to price decreases – very much like web-scraped data –, the troughs are highly correlated to sale periods.
- How much (more) information is contained in price indices based on scanner or web-scraped data compared to traditional methods? *Volatility* and *bias* are some *further issues* in this field.
- Indices from scanner and web-scraped data have shown to be more volatile than traditional indices. While the traditional price collection of matched models shows little to no noise in the price developments, the new methods introduce a lot of noise in the time series. This is all the more true for weighted indices and using scanner data.

Bias

- Scanner and web-scraped data represent an admittedly “big” but biased non-probabilistic sample – not the population. There are transactions that are in the scope but are not recorded electronically, not available to the statistical office, deleted in the filtering step, cannot be matched or linked, and so forth. After all, not more data are better, better data are better.
- Scanner and web-scraped data can be very precise but at the same time may have limited accuracy. The danger lies in blindly trusting that these new data sources must give us better answers; in fact, big data are not capturing all transactions, just some, and we might not even know which ones are missing. That is why the combination of more traditional data with big data is the ticket to reducing coverage bias.
- Scanner and web-scraped data are no panacea!

3. The use of big data sources for consumer price indices

- Processing data
- Classification
- Index calculation
- Further issues

Appendix 1. The European Harmonised Index of Consumer Prices

- Numerical example

Aggregation of chain-linked sub-indices

- From time to time, it may be necessary to produce bespoke aggregates which are often requested by users.
- It is important to note that only *unchained indices* should be aggregated. This applies to all levels of index aggregation. Once chain-linked, index number series are no longer *consistent in aggregation*.
- To obtain unchained index numbers from their chain-linked counterparts, compilers must divide the chain-linked index number of each month of each year by the chain-linked December index of the previous year (and multiply by 100). The unchained index for month of year is calculated by solving for the short-term series:

$$P^{0t,mt} = \frac{CP^{b,mt}}{CP^{b,12(t-1)}}.$$

Numerical example

- In order to produce bespoke aggregates from published chain-linked index numbers, the first step is always to unchain the relevant components; the starting point is an *unchained index number series*. The unchained aggregate index numbers are then aggregated together using their relevant weights to produce unchained index number series for the new bespoke aggregate. These index numbers are then chain-linked.
- The Excel file available on CIRCABC contains the data referring to the changing composition of the euro area are those published by Eurostat on 22 February 2017. The *indices* and *item weights* (blue tabs) are both rounded to two decimals, which will affect the calculation results. They are, nonetheless, exact except for rounding differences.
- The five special aggregates – processed food, unprocessed food, non-energy industrial goods, energy and services – are *aggregated* (grey tab) to the total HICP.

Excel file

Harmonised Index of Consumer Prices ([1]) - Aggregation
(Last update: 22 February 2017)

	Unchaining, Equation (8.21)			Item weights	Aggregation, Equation (8.18)	Chain-linking, Equation (8.20)	Index, 2015=100	
	Processed food	…	Services	…	Total	Total	Total	Zero check
2013M12	:	:	:	:	:	100.11	:	:
2014M01	100.32	…	99.60	…	98.88	98.99	98.99	0.00
2014M02	100.40	…	100.11	…	99.19	99.30	99.30	0.00
2014M03	100.48	…	100.31	…	100.12	100.23	100.23	0.00
2014M04	100.45	…	100.44	…	100.27	100.38	100.38	0.00
2014M05	100.53	…	100.28	…	100.16	100.27	100.27	0.00
2014M06	100.45	…	100.73	…	100.27	100.38	100.38	0.00
2014M07	100.59	…	101.67	…	99.62	99.73	99.72	0.01
2014M08	100.61	…	101.97	…	99.73	99.84	99.84	0.00
2014M09	100.62	…	100.89	…	100.17	100.28	100.28	0.00
2014M10	100.67	…	100.68	…	100.11	100.22	100.22	0.00
2014M11	100.60	…	100.52	…	99.93	100.04	100.04	0.00
2014M12	100.55	…	101.24	…	99.83	99.94	99.94	0.00
2015M01	100.18	…	99.40	…	98.45	98.40	98.40	0.00
2015M02	100.39	…	100.06	…	99.09	99.03	99.03	0.00
2015M03	100.53	…	100.08	…	100.20	100.14	100.15	-0.01
2015M04	100.62	…	100.15	…	100.45	100.39	100.39	0.00
2015M05	100.55	…	100.36	…	100.67	100.61	100.61	0.00
2015M06	100.60	…	100.58	…	100.66	100.60	100.60	0.00
2015M07	100.66	…	101.65	…	100.02	99.96	99.96	0.00
2015M08	100.70	…	101.96	…	100.03	99.97	99.97	0.00
2015M09	100.64	…	100.89	…	100.25	100.19	100.19	0.00
2015M10	100.70	…	100.78	…	100.39	100.33	100.34	-0.01
2015M11	100.78	…	100.43	…	100.24	100.19	100.19	0.00
2015M12	100.72	…	101.13	…	100.23	100.18	100.17	0.01
2016M01	100.22	…	99.42	…	98.55	98.72	98.72	0.00
2016M02	100.25	…	99.85	…	98.71	98.88	98.88	0.00
2016M03	100.24	…	100.32	…	99.93	100.11	100.11	0.00
2016M04	100.43	…	99.91	…	99.97	100.15	100.15	0.00
2016M05	100.39	…	100.25	…	100.34	100.51	100.51	0.00
2016M06	100.40	…	100.59	…	100.50	100.68	100.68	0.00
2016M07	100.41	…	101.69	…	99.95	100.13	100.12	0.01
2016M08	100.46	…	101.93	…	100.03	100.20	100.21	-0.01
2016M09	100.40	…	100.91	…	100.42	100.60	100.60	0.00
2016M10	100.49	…	100.73	…	100.68	100.85	100.85	0.00
2016M11	100.77	…	100.39	…	100.58	100.76	100.76	0.00
2016M12	100.73	…	101.25	…	101.12	101.30	101.31	-0.01

([1]) Data refer to the changing composition of the euro area.
Source: Eurostat