Kristiina Nieminen, Antti Suoperä, Satu Montonen,
Hannele Markkanen

5/7/2024

The Pyramid of Decisions --Construction of Bilateral Index Using Alternative New Data Sources

The aim of presentation

The increasing stream of digital data has opened up new possibilities for national statistical organizations. New alternative data sources (ADS) like transaction data, web-scraped data, and administrative data can improve quality and coverage of Consumer Price index while minimizing costs and burden of data collection.

With the new data, choosing the calculation method is a challenge for CPI-expert. Which index calculation method should we apply and what other decisions should we make when utilizing new data sources in the CPI production?

In this presentation, we break-down the problem into smaller components, so called decision-making-points. The core of the presentation is the pyramid of decisions, that presents alternatives as a continuum, starting from microdata and moving upwards level by level replying to questions on decision points.

The decision pyramid presents undetermined number of potential paths of decisions that varies according to data source. Simple path of decisions may be more pronounced for web-scraped data while complete scanner-data enable use of more sophisticated methods.

Background

In Finland, the scanner data was first tested in the year 2000 when the hedonic method was investigated for electronic goods. However, the more intensive testing began during the period **2015-2016** in the EU-funded project. The first complete dataset consisted of pharmaceuticals, providing average prices, sales quantities, and extensive registry information for categorizing products to COICOP-classification.

At first testing covered basic formulas (e.g Laspeyres, Paasche) and excellent index number formulas such as, Fisher, Törnqvist, Stuvel, Sato-Vartia. This approach was chosen as observed data included large deviations in quantities and/or prices, and naturally large variation of value shares.

We found out that contingently biased indices, like simple base or observation period weighted indices (L and Pa), get values here and there, but superlative indices (based on symmetrical weights) go hand in hand. Choosing among the superlative or excellent index number formulas was more or less indifferent, because their values differ so little.

In **2017**, alternative strategies for creating the index were tested with pharmaceutical data, as it was known that the chained strategies may be biased compared with the base strategies. Thus, tested strategies were base (=fixed base), chain in isolation (fixed commodity basket), pure chain (new and discontinued products taken into account) and mixed strategy. The mixed strategy was combination of base strategy for "old" products supplemented with chained strategy for "new" products. The fixed base method gave most reliable results, but its disadvantage was the delay in the inclusion of new products until the next update of the base period.

Thereafter, testing was continued with new scanner-data of daily products sales. In **2018**, the GEKS-method and magnitude of chain error was tested using multi-period identity test (MPIT) and path-independence-test (PIT). Our results presented empirically that GEKS index constructed by chaining windows together leads to serious chain error for the commodities having largely varying consumption within a year. Transitivity holds only in fixed window.

The index number theory based on bilateral or multilateral methods is not possible for dwellings using matched pairs because transacted dwellings appear mostly only once in data. Therefore, hedonic method was tested in **2019** for dwellings. The used data was high-quality register data on all free-market transactions of dwellings in new apartments and terraced houses. Data included statistical unit specific information of unit prices, quantities, values and some quality characteristics (region, type of apartment, square meters, distance from the center of municipal services, owned or rented). In the hedonic price decomposition average price change was divided into quality corrections and quality adjusted price change for all averages separately. This was done by using the Blinder-Oaxaca decomposition.

In our previous tests, we discovered that quantities consumed can vary significantly, sometimes even tenfold, between different time periods. This variation poses a challenge when estimating price changes based on subgroup averages. The bias introduced by such substantial changes in quantities can significantly impact the accuracy of our calculations. To mitigate this bias, one approach is to consider hedonic methods.

This was the reasoning for studies made in 2020 for alcoholic beverages. All index number formulas and basic averages were presented by their logarithmic representations that made it possible to compare these statistics pair wisely and understand relationship. Differences of two price change were investigated and as a result it was shown that when quantities change substantially and not proportionally, it introduces a significant bias in the index calculations. Therefore, a hedonic method was tested to remove quantity changes from the actual log-price change separately for the basic averages.

Our latest studies (in 2021-2024) covered the problem of continually declining prices, like clothes, shoes, mobile phones, computers, and hedonic methods applied for second hand cars and housing prices.

The price index for used cars (published in 2023) relies on a stratified approach. The hedonic approach involves employing price modelling with detailed stratification and heterogenous behaviours. Within each stratum, logarithmic price ratios are computed and furthermore, utilize the Blinder-Oaxaca decomposition method, which allows us to break down the actual (unbiased arithmetic or geometric) average price ratios into two components: quality correction and quality-adjusted price ratios, all in logarithmic form. It offers an explicit interpretation of quality adjustment, encompassing both quality corrections and quality-adjusted price changes for all strata.

The price index for block of flats (=apartments) and terraced houses (published in 2023). This study provides an alternative hedonic solution for quality adjusting than hedonic time-dummy approach. The study give lessons about how ‘weighted-by-economic-importance’ should be done using transparent algebra of unbiased estimates. The findings and

suggestions were following: 1. Use detailed partitions that are both based on postal code areas and smaller municipalities, and room numbers in estimation of price models. 2. Form price decompositions for stratums and aggregate them into ‘crude’ levels using excellent formulas, say for example Törnqvist. 3. Use aggregation rule of weighted arithmetic average, because of standard practice of publishing average prices.

Alternative data sources for the CPI

The Consumer Price Index (CPI) utilizes data from various data sources and collected in various ways because a single collection method cannot cover the entire basket of goods’ information needs. Additionally, the main goal of the organization often involves streamlining and automating data collection to minimize costs and reporting burden.

Traditionally, the price data was collected by price interviewer-conducted visits and by CPI-team (a.k.a direct data collection). However, since the 2000s, there has been an increasing trend toward utilizing electronically available datasets. These new alternative data sources include:

- Administrative Registers
- Retail Sales Data (so scanner Data): Captures actual sales transactions.
- Web-Scraped Price Information: Extracted from company websites.
- API-Driven Data Retrieval: Obtained via programming interfaces.

The table below presents alternative data sources from five perspectives: Acquisition, Coverage, Product Identification, Completeness of data and Timing.

DATA DIMENSION	TRADITIONAL LOCAL COLLECTION BY PRICE INTERVIEWERS	DIRECT DATA COLLECTION BY CPI-TEAM	SCANNER-DATA	WEB-SCRAPED DATA	ADMINISTRATIVE DATA AND REGISTERS
Data acquisition	Manual	Manual	Automatic	Automatic	Automatic
Coverage	Sample of items and retailers	Sample of items and retailers	All transactions	Bulk of products or sample of products	Census or almost census
	One or few observations	One or few observations	All or sample of retailers	Sample of online retailers	
Product identification	Product name, item description, some characteristics	Product name, item description, some characteristics	Product ID, item description, characteristics	Product ID, item description, characteristics and attributes	Elementary units idetified by key-codes
			Retailer specific product categories	Retailer specific product categories	Prices, metadata and some cases quantities
Completeness of data	Only shelf-prices	Prices, sometimes also quantities	Sale value and sold quantities OR prices and quantities	Prices and metadata on products. No quantities	Prices and metadata on products. No quantities
Timing	Single collection day	Single collection day or whole month	Transactions aggregated by day, week or month	Single collection day, several times a month	Month

Data acquisition involves either manual work conducted by a price-interviewer or a CPI-expert. Alternatively, it may consist of automatic file transfers from the data supplier to the statistical office or procedures

implemented by the statistical office to collect data from websites or interfaces provided by companies.

The coverage of data varies depending on the data source. The data may consist traditional sample-based data collection, where the sample typically covers the most representative products and services. Alternatively, the data may encompass a company's entire monthly sales (scanner-data), include the complete registry of an administrative entity (administrative data), or represent a cross-section of prices for available products and services (web-scraped data).

Product Identification: Traditionally, products and services have been identified based on the product descriptions, the quality definitions, and the associated commodity classification. In new datasets, individual products or services often utilize unique identification codes that allow seamless identification of a single item from a large set of commodities. Additionally, these datasets include at least the name of the product/service, and perhaps even detailed descriptions and/or feature information.

Completeness of data Traditionally price data covered consumer prices collected from the edge of the shelf, without accounting for discounts. With the advent of new datasets, it became possible to access actual sales and sold quantities (scanner-data) or usage (e.g., rentals) from administrative data. The web-scraped data contains metadata, detailed information about product characteristics, in addition to prices.

Temporal Dimension: Traditional data collection typically focused on shelf-price or list price, that usually represents prices for a single day. In contrast, scanner data includes all transactions, prices, and quantities, aggregated to daily/weekly/monthly levels to minimize size of dataset. This data also includes discounts provided to consumers. On the other hand, web prices represent a cross-section from the specific collection time, but this temporal dimension can be improved by implementing multiple automated selections, such as weekly snapshots.

Alternative index number formulas and construction strategies

The objective of the index calculation is to provide high-quality, comparable measures of consumer price inflation (Eurostat, 2024). To achieve this objective, it is crucial to ensure compliance with criteria, particularly when transitioning from traditional data collection practices to enhanced procedures that utilize big data on a large scale.

Latest updated manuals: CPI concepts and methods 2020 and HICP methodological manual 2024 (Eurostat) provide a list of suitable formulas for the compilation of CPI. These are included in the table below.

The CPI (Consumer Price Index) and the HICP (Harmonised Indices of Consumer Prices) both apply Laspeyres-type index formula at the level of elementary price index and above as weights are available. The weight reference period is calendar year, or 12 consecutive months, and the price reference period is a period which prices are used as denominators in the index calculation (e.g in the HICP it's December of previous year).

At the elementary aggregate following averages can be used:

- Jevons, that is the ratio of the geometric mean of prices or the geometric mean of the ratio of prices

- Dutot, that is ratio of arithmetic mean of prices
- Carli, that is arithmetic mean of price relatives

An elementary price index is typically calculated from two sets of matched price observations.

With the advent of new scanner data, NSIs gradually transit to using superlative formulas that leverage current and base period consumption shares, thus better accounting for changes in consumption occurred previously.

The table below summarises most known index number formulas and presents key features for all of them.

	Old weights	New weights	Bias**	Consistency in aggregation	Time reversal test	Transitivity test	Commensurability test	Proportionality test	Factor reversal test	Determinatene ss test	Fisher proportionality test ****
Laspeyres (La)	X		CB	X	X		X	X		X	X
Geom. Laspeyres	X		CB	X	X		X	X			X
Harm. Laspeyres	X		CB	X	X		X	X			X
Palgrave (Pl)		X	CB	X	X		X	X		X	X
Geom. Paasche		X	CB	X	X		X	X			X
Paasche		X	CB	X	X		X	X		X	X
Jevons			CB	X	X	X	X	X			X
Dutot			CB	X	X	X		X			X
Carli			CB	X			X	X		X	X
Fisher	X	X	0		X		X	X	X	X	X
Stuvel	X	X	0	X	X		X	X	X	X	X
Törnqvist	X	X	0		X		X	X			X
Sato-Vartia	X	X	0		X		X	X	X	X	X
Marshall-Edgeworth	X	X	0	X	X		X	X		X	X
Walsh	X	X	0	X	X		X	X		X	*

- First set of formulas are basic formulas (old vs. new weights)
  - Laspeyres, Geometric Laspeyres, Harmonic Laspeyres
  - Palgrave, Geometric Paasche, Paasche
- Second set of formulas are averages Jevons, Dutot and Carli that are employed when item-specific weight information is unavailable.
- Third set of formulas are superlative or other type index number formulas that are considered excellent because their bias is zero.
  - Fisher, Stuvel, Törnqvist,

In the table, the first set of columns (1-4) describes features of the index number formulas, while the second set (5-11) outlines properties that different kind of indices compiled using specific formula.

Weights: The basic formulas use either old or new weights, while excellent formulas use the weight information from both periods, base and comparison period.

The **Bias** column highlights index number formulas that are either contingently biased (CB), permanently biased (PB) or give accurate results for small changes. The size of bias depends on the data in

question: it may be small or large. Those formulas that present zero (0) in the column are unbiased all the time.

Consistency in aggregation-column highlights formulas that meet the requirement for exact consistency in aggregation. Consistency in aggregation means that if an index is calculated stepwise by aggregating lower-level indices to obtain indices at progressively higher levels of aggregation, the same overall result should be obtained as if the calculation had been made in one step (CPI manual 2020, p182). For example, Fisher and Törnqvist index number formulas are not exactly consistent in aggregation, instead they are approximately consistent in aggregation.

According to the international CPI and HICP manuals one way to decide upon an appropriate index formula is to require it to satisfy certain specified axioms or tests. **Four basic tests** are emphasized to illustrate the axiomatic approach:

- *Time reversal test*: The index from period 0 to period t should equal the reciprocal of the index from t to 0
- *Transitivity test*: The index from 0 to 1 multiplied (chained) by an index from 1 to 2 should equal a direct index from 0 to 2
- *Commensurability test:* The index should be invariant compared to the unit in which prices are recorded
- *Proportionality test:* If all prices change x%, the index should also change by x%. A special case of this test is the *identity test*, which requires that if the price of every variety is the same as in the reference period, the index should be equal to unity

Following tests should be also acknowledged when analysing appropriate index formula:

- *Factor reversal test:* the product of the price index and the quantity index calculated by same index number formula ought to equal the value ratio
- *Determinateness test:* means that, if any single price p^k or quantity $q^k, k = 0,1$ tends to zero, then the price index should not tend to zero or infinity
- *The Fisher proportionality test:* assesses whether a price index remains unchanged when the units of measurement for each commodity change. In other words, it examines whether altering the units of measurement affects the index value.

Index compilation methods

The international manuals provide (in principle) two methods in the index number theory for index compilation: *the base* and *the chain method*.

The *base method* uses binary comparisons from base period 0 to observation period t and the chain method from (t−1) to t,t=1,2,… . Both methods have weak points. The weakness of base method is that it excludes new and disappearing commodities from index calculations.

On contrary, the *chained* method may suffer from chain drift or chain link bias. Chain drift occurs if a chained index “does not return to unity when prices in the current period return to their levels in the base period” (ILO,

2004, p. 445). When there are large period-to-period fluctuations in prices, quantities and values, all kinds of chained price indices should be avoided, because the chain drift may occur. The base method never suffers from the chain drift whatever the index number formula is.

In addition to these two, the method for index compilation could be also a *mixed method* which is combination of the base method for evenly consumed products and the chained method for new and disappearing commodities. These two methods, or more precisely their price changes, could be weighted together by the value shares and then used in the construction of index series.

The pyramid of decision

When incorporating new data sources into compilation of the Consumer Price Index (CPI), index compilers must consider several crucial aspects beyond the formula and index compilation method. Let's explore these additional considerations in the pyramid of decisions below. The levels of the pyramid are categorised into four groups:

1. Data quality and reliability
    a. Ensure that each product or service in the data can be uniquely identified. This is crucial for accurate tracking and comparisons.
    b. Validate the data to ensure its accuracy, consistency, and reliability. Consider supplementing it with additional information, such as product characteristics and weights.
2. Coverage of data
    a. Assess whether the data source covers all relevant prices and determine if any observations need to be filtered out, for example excluding outliers or irrelevant items.
3. Special cases
    a. Define consistent approach for handling missing prices. Consider imputation methods or replacement of the old product with new product.
    b. Address how to handle new products entering the market or existing products disappearing. Use quality adjustment methods when justified.
4. Data aggregation and weights
    a. Decide how to aggregate the data. Each decision has implications for the resulting index.
    b. Ensure that the weights assigned to different items or homogenous products accurately reflect their importance in consumer expenditures.

Kristiina Nieminen, Antti Suoperä, Satu Montonen,
Hannele Markkanen 5/7/2024

Detailed examples of the aspects are given in the appendix 1 and four possible paths for ADS-datasets in the appendix 2

Incorporating new data sources in the CPI production is a delicate process, and thorough evaluation and validation are essential to maintaining the CPI's accuracy and relevance.

Conclusions

New data sources create new opportunities for evaluating development of consumer prices, but also pose challenges for the expertise of statisticians and the IT infrastructure used for data processing. New data transfer processes need to be established for data reception, and validation methods for the data must be developed, as the size of data can be up to 10-100 times larger than before.

These new data sources have been tested from various perspectives for the past 9 years at Statistics Finland. The testing has focused on pairwise comparisons (i.e matching-pairs) following the principles of the consumer price index.

Using unique identification codes such as GTIN-codes, products can be identified without gaps, making this method suitable when consumption remains consistent from month to month and year to year. Individual products may be grouped into so-called homogeneous product categories (i.e HPs) when attrition rate is high, product consumption varies significantly between months, or when a single product or service appears only once in the dataset (e.g. new dwellings).

Some of the traditional practices have been dropped, e.g imputation and replacement-procedures when the dynamic basket approach has been introduced. The benefits and drawbacks of alternative index formulas and compilation strategies have been tested.

Based on these experiences and findings, the table of index number formulas has been updated and a pyramid of decisions compiled. Going

through these components will help to identify, evaluate, and make decisions regarding alternative index compilation methods.

However, the final choice is particularly influenced by the provided data:

- Coverage and content
- Quality and reliability
- Characteristics of the products and quality changes
- Availability of weight information.

Following criterion have been considered when choosing an appropriate index method:

1. Compliance with international standards,
2. Simplicity in the production and interpretability of the results
3. No chain drift. Thus, select base strategy.
4. The basic averages are good price change estimates only if quantities have not changed or have changed proportionally.
5. A hedonic method is useful when bilateral methods (i.e matching of products) may not be used.
6. Any index number formula is suitable if products are sold evenly from month-to-month. It is the strategy that matters more in these cases.
7. Any superlative or excellent index number formula is applicable when complete data on sales value and quantities is available.

Kristiina Nieminen, Antti Suoperä, Satu Montonen,
Hannele Markkanen

5/7/2024

Appendices

Appendix 1: Examples of the aspects in the pyramid of decisions

Data Quality and Reliability

Product Identification:

Alternative methods for product identification are either to use individual identification codes or Homogeneous Products, HP.

- GTIN Codes are global trade item numbers that precisely identify products and services.
- Stock Keeping Unit (SKU) Codes, SKUs, help to track inventory automatically in retail stores. They consist of alphanumeric digits and a scannable barcode, providing details on price, product, manufacturer, and point-of-sale. Some retailers may provide these in addition to GTIN. One SKU includes one or more GTIN codes.
- Price-Look-Up (PLU) Codes are mainly used for Fresh Products: These are slightly broader four-digit codes that eliminate the need for visual identification during checkout and inventory control. PLUs are crucial when similar-looking products or items (e.g., organic and conventional varieties) have different prices.

These Identification Codes facilitate easy pairwise comparisons, so called matching-pairs. However, when using a fixed base method, new products are included in the index only after the annual base period update. Alternatively, Homogeneous Products (HP) can be defined based on common characteristics. The average price of products within a group is calculated. This approach allows new products to be incorporated into the index promptly. For instance, clothing can be grouped by brand and other attributes alongside the COICOP classification.

Pre-processing of data

To ensure quality of the price observations and inclusion in the index compilation, collected data need some pre-processing that is dependent on the nature of the data source. Traditionally collected data is typically validated immediately as CPI-team receives the data. All other data sources need to be mapped to COICOP-categories, the naming of variables should be standardised and maybe some supplementary data could be added.

Coverage of data

Data Filtering:

First, products that cannot be uniquely identified or do not belong to the selected commodity groups are excluded. Then the need for other filters is assessed. When using the Jevons formula, filters are useful to avoid extreme cases that might distort price development., for example consider outliers in price observations—cases with exceptionally high/low sales values or prices (outliers), very low sales values (low sales), or significant price changes (dump filter).

Scanner-data is so called complete data having detailed information on monthly consumption and turnover. In theory, filters should not be

necessary. However, Finland has chosen to exclude very large price changes from index calculations (price ratio >XX) even in this situation.

Sample Selection:

For new data sources, it is good to follow the standard guidelines. Create a outlet sample (select only a necessary set of individual companies or retailers). Based on this sample, choose companies and their products and services for data collection and computation.

Alternatively, aim to geographically cover the entire country's sales. This choice depends on the data supplier's willingness to cooperate and to provide enough new data. For ensuring coverage, including chain stores operating in various regions (of different sizes and types) in the sample would be beneficial.

For web-scraped data product sampling is crucial. The first step is to define the outlet sample, then consider the product sample. Decide whether to include samples of products (concentrate on representative products similarly to traditional price collection) or whether to focus on representing as many products as possible in index compilation.

Special Cases

Treatment of Missing Prices:

In traditional price collection, temporarily missing prices are imputed using price changes observed in other items within the same COICOP group or sub-class. Permanently discontinued products/services are replaced with new ones. This ensures that the commodity basket remain consistent throughout the year.

Scanner data provides a significantly larger number of price observations compared to traditional methods. Therefore, replacement procedures may be too laborious to carry out. Also, scanner-data may not require imputation of missing prices due to abundance of price observations. Instead, the dynamic basket approach may be used to adjust the basket of goods and services over time based on actual consumption patterns. The dynamic basket approach, coupled with appropriate index number formulas, ensures flexibility and accuracy of CPI.

Treatment of Seasonal Products

Traditionally seasonal products receive special treatment compared to regular items. This distinction arises because seasonal products are not available every month on the market, instead availability follows some cyclical annual pattern. For example, strawberries and cherries are available in minimal quantities during the out-of-season period (say winter months), thus prices are not collected during the winter-season. These missing prices need to be imputed each month during the winter-season.

The HICP manual offers two methods for imputation of missing season products: Seasonal Imputation Method (targeted or all-product imputation method) and Variable Weights Method.

In scanner data, missing prices can be imputed following the guidelines above. Alternatively, a dynamic commodity basket can be used, ensuring that seasonal products are part of the base period's product set.

Quality adjustment

Quality adjustment is needed for when old product (the replaced product) is replaced with the new product (the replacement product) and when replacement product is not comparable with replaced product. The implicit or explicit quality adjustment methods may be applied for non-comparable replacement product. See CPI manual p.132-148 / HICP manual 6.5.1 Definitions p. 174 (annex 3)

The Oaxaca-Blinder decompostion: also known as Kitagawa decomposition, is a statistical method that explains the difference in the means of a dependent variable between two groups by decomposing the gap into within-group and between-group differences in the effect of the explanatory variable (Wikipedia). In Finnish CPI it is used by.breaking the change of average prices (i.e. arithmetic and geometric) down into quality adjustment factors and price change standardized for quality.

Data Aggregation and Weighting

Definition of the Commodity Basket:

The commodity basket can be defined as fixed for individually included products, meaning that the basket remains the same throughout the year. Alternatively, it can be dynamic/flexible, considering only the products present in both the current and base periods.

The latter method is particularly suitable for scanner data and potentially for administrative data, provided that the dataset size (number of observations) is sufficient.

Weights:

The possibilities for determining product- or service-specific weights depend significantly on the collected data.

In traditional price collection, company-specific consumption shares are typically not available. Therefore, the only way to manage internal weights is to define the desired number of price observations per company, implicitly ensuring the correct weighting of companies. Neither the item-specific weights are used as the goal is to collect prices of similar products under one commodity based on quality definitions (=product description).

For web-scraped data, annual sales values, the weights, can be collected through other means (e.g., directly from companies). This is potential approach for weighting these prices as some companies are more willing to provide annual sales data by products instead of monthly detailed data.

Scanner data and potentially administrative data allow for considering consumption shares of two periods, current and base period.

Definition of Base Period in the Fixed Base Method:

For scanner data, the base period can be defined in various ways because the dataset is considered “complete,” containing sales values and quantities for all periods. The base period options include:

- Previous Year’s December: For instance, in Europe, weights are updated to December to align these with the price reference period that is December t-1. However, it’s important to exercise caution when using December as the base period because consumption in December significantly differs from other months.

- Base Year of the Index or a Specific Month from That Year: This approach is no more used in Europe due to the transition to annually chained index.
- Monthly Average Values from the Previous Year (the “Finnish Method”): The advantage is that the commodity basket includes all products sold during the previous year, including all seasonal items, regardless of whether they were sold in multiple months or only one month.

Appendix 2: Examples of possible index compilation paths

Traditional data

- *Data:* traditionally collected price observations for durables and services
- *Identification of product:* Product description and -specification, name of the product and some price determining characteristics
- *Pre-processing of data:* All price observations collected by price-interviewers are validated as they are transferred to database. Based on the checking lists, CPI-experts check small set of the observation manually. No linking to COICOP needed as these products are linked to COICOP in the beginning of the year. No need for other proposed tasks either.
- *Filtering of data:* No filtering done. All price observations need to be taken in the index compilation.
- *Sampling:* The sampling of outlets and products done on annual basis according to the HICP-manual.
- *Treatment of missing prices:* The temporarily missing prices are imputed based on the price change (of the products belonging to same region and 7-digit coicop-category) from the preceding to the present period.
- *Treatment of Quality Changes:* The replacement of permanently missing products done at the third month of disappearance. Comparability of replacement product and replaced product assessed currently manually based on the product characteristics.
- *Treatment of Seasonal Products:* Seasonal weights method applied. This means that infra-annual (i.e monthly) weights are defined for all commodities belonging to the basket. Weight of seasonal product is set to zero (=) only when the product in question is expected to be out of season and the weight is reallocated to neighbouring 6-digit sub-sub-classes.
- *Basket:* Fixed basket
- *Weights:* Annual weights
- *Index Compilation Method:* Fixed base method
- *Definition of Base Period:* December t-1
- *Index number formula:* Jevons below the EA, Laspeyres-type index from the lower-level index (=EA) to higher level COICOP categories

- *Result:* The microindex for 6 regions broken-down to 7-digit COICOPs

Scanner-data: **Alcoholic beverages**

- *Data:* monthly scanner-data on alcoholic beverages
- *Identification of product:* Company specific individual identification codes
- *Pre-processing of data:* Standardise variable names and types. No need to derive unit-prices as data contains them. Individual products are linked to COICOP categories 02.1.
- *Filtering of data: U*nknown products, incomplete products and extreme price changes are filtered out.
- *Sampling:* No sampling carried out. All products are taken. Company follows national pricing strategy, meaning that same prices are applied in each region and outlet.
- *Treatment of missing prices:* No imputation, no replacement
- *Treatment of Quality Changes:* No quality issue as matching of products is done with the unique identification code.
- *Treatment of Seasonal Products:* Seasonal products are acknowledged correctly as those are included in the basket.
- *Basket:* Dynamic basket that is updated annually
- *Definition of Base Period:* An average month of previous year
- *Weights:* An average of current and base period weights
- *Index Compilation Method:* Fixed base
- *Index number formula:* Törnqvist
- *Result:* Microindex of alcoholic beverages for state-owned alcohol monopoly

Web-scrape data: Holiday cottages

- *Identification of product:* Company specific individual identification code for each holiday cottage
- *Pre-processing of data:* Standardise variable names and types and determine the time periods. Individual products are linked to one COICOP category in 11.2.0.2 since all holiday cottages belong to the same COICOP.
- *Filtering of data:* No filtering done. All price observations need to be taken in the index compilation.
- *Sampling:* The sampling of holiday cottages is done on annual basis.
- *Treatment of missing prices:* The temporarily missing prices are imputed based on the price change (of the products belonging to 7-digit coicop-category) from the preceding to the present period.
- *Treatment of Quality Changes:* No quality issue as matching of products is done with the unique identification code.

- *Treatment of Seasonal Products:* Seasonality is not taken into account since there are always holiday cottages available for reservations.
- *Basket:* Fixed base method
- *Definition of Base Period:* December t-1
- *Weights:* No weights are available.
- *Index Compilation Method:* Fixed base method
- *Index number formula:* Jevons
- *Result:* Microindex for holiday cottages

Administrative data: In this example we use results of the study "Hedonic Price Index Number for New Blocks of Flats and Terraced Houses in Finland, 2019". We start from the bottom of the pyramid and select one or several options at each level (= question)

- *Data:* quarterly administrative register data
- *Identification of product:* Create homogenous products by dividing twelve (12) regions to four types of apartments (one-room, two-rooms, three-rooms or bigger in block of flats and terraced houses)
- *Pre-processing of data:* Supplement register data with additional information about the apartment characteristics. Define stratums for hedonic model by using parameters: *location, square meters, flat type, distance to municipal center, owned/rented* and specify the price model for Homogenous Products
- *Filtering of data:* Take all valid price observation. Exclude outliers, extreme prices and unknown apartments.
- *Sampling:* No sampling. Take all.
- *Treatment of missing **information**:* Missing variables are derived based on other available information.
- *Treatment of Quality Changes:* the quality differences from price index numbers based on unweighted or weighted arithmetic and geometric averages need to be removed. For controlling the quality change of characteristics, we used regression analysis, i.e hedonics.
- *Treatment of Seasonal Products:* not relevant
- *Basket:* dynamic basket.
- *Definition of Base Period:* an average quarter of previous year
- *Weights:* an average of current and base period weights
- *Index Compilation Method:* fixed base method
- *Index number formula:* Törnqvist
- *Result:* The price index of new apartments

Appendix 3: Implicit and explicit imputation methods

Implicit methods (HICP manual, table 6.7.14):

Method and Source of price change

- *Bridged overlap*: Individual product in the same elementary aggregate (EA)
- *Overall mean imputation*: All individual products in the same EA.
- Targeted mean imputation: Selected individual products in the same EA.
- *Class mean imputation*: Quality adjusted replacement individual products in the same EA.
- *Monthly chaining and replenishment* (MCR): All individual products in the same EA.
- *Backcasting* (base price imputation): All individual products in the same EA, price change since December.
- *Link-to-show-no-price change*: Selected individual products in the same EA.

Explicit methods (HICP manual, 6.5.1 Definitions)

Method and Source of price change

- *Direct price comparison*: no quality difference
- *Package-size adjustment:* the value of a change in package size, as a proportion of the price, is assessed as the relative change in package size.
- *Single-variable adjustment*: the value of the quality change between a replaced and a replacement individual product, as a proportion of the price, is assessed as the relative change in some function of one characteristic of the individual product.
- *Option pricing*: the value of the quality change between a replaced and a replacement individual product is assessed as some fixed proportion of the market price of features by which the two individual products differ.
- *Supported judgement*: the value of the quality change between a replaced and a replacement individual product is calculated by using supplementary information sources.
- *Hedonic regression method*: the quality adjustment is in some way based on a regression equation, which expresses the price as a function of product characteristics. This model estimates coefficients for each product characteristics which are proxies of the implicit prices of these characteristics, often referred to as “shadow prices”.
- *Combined quality adjustment method*: the value of the quality change between a replaced and a replacement individual product is assessed using a combination of methods.

Kristiina Nieminen, Antti Suoperä, Satu Montonen,
Hannele Markkanen

5/7/2024

Excellent formulas

Fisher price index: The geometric average of the Laspeyres price index and the Paasche price index. It is a symmetric index and a superlative index.

Törnqvist price index: symmetric index defined as the weighted geometric average of the price relatives in which the weights are simple arithmetic averages of the expenditure shares in the two periods. It is a superlative index. Also known as the Törnqvist–Theil price index.

Walsh price index A basket index in which the quantities are geometric averages of the quantities in the two periods. It is a symmetric index and a superlative index.