Integration of new data sources along traditional collection

Claude Lamboray | EUROSTAT C4 | 18/09/2020

Aggregation structures in a CPI

- CPIs are organized according to **hierarchal classifications (e.g. ECOICOP)**.
- Product categories (‘items’) may be further **stratified according to region and type of outlet**.
- The **elementary aggregate** is the smallest aggregate for which expenditure weights are used.
- Within an elementary aggregate, outlets are sampled and **product-offers are selected for pricing** in these outlets.
- Elementary indices are obtained using an unweighted index formula (e.g. **Jevons index**) and aggregated to higher-levels using the expenditure weights (e.g. **Laspeyres-type index**).

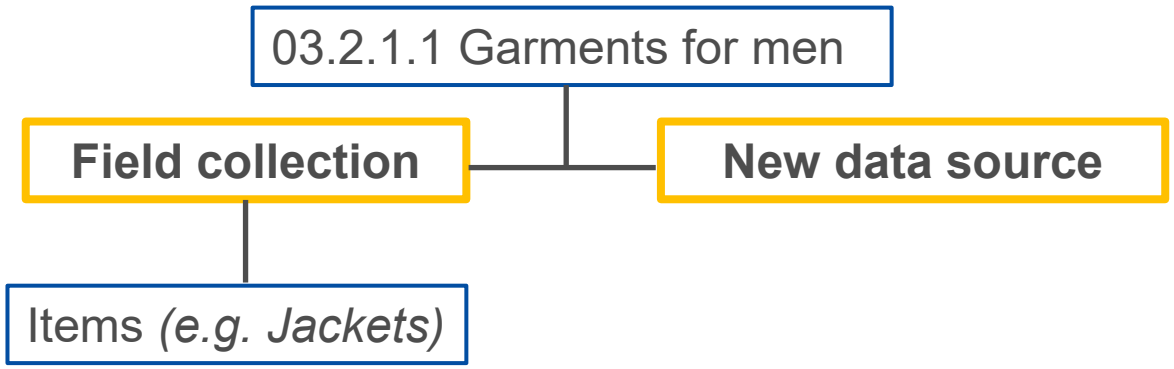
Aggregation structures in a CPI

- Treat the prices obtained from new data sources the same way as prices collected in the field and **keep the methodology largely unchanged**.
- **Adjust the standard two-stage aggregation system** to the new data sources because:
  - The new data sources have a wider coverage than prices collected in the field.
  - The new data sources are typically characterized by being dynamic.
  - Weights may (or may not) be available at more detailed levels.

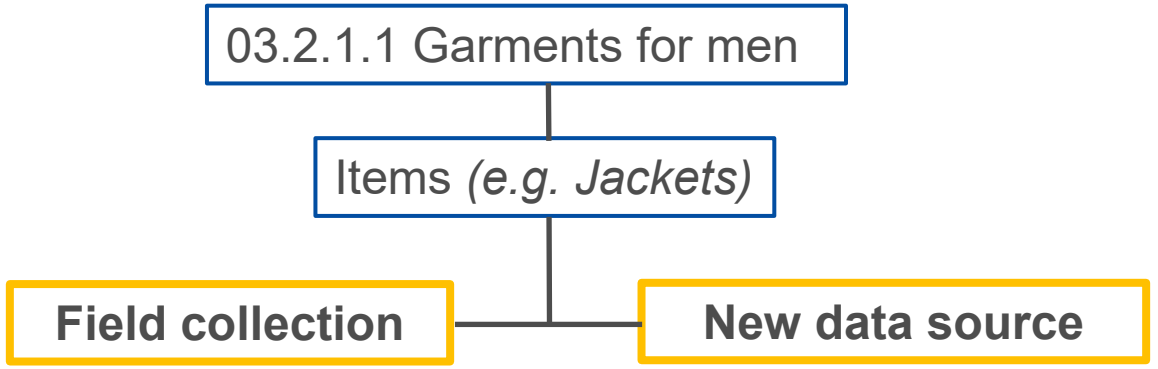
→ Data can be stratified/structured in different ways and there is a large range of possible index compilation methods.

Data integration

At which level should the new data source be introduced?



Integration at ECOICOP subclass level



Integration at item level

European Commission

Data integration

- Each **data source** corresponds to a **stratum** to which a **fixed weight** must be attached and for which a **price index** must be estimated.
- **What do the weights represent?** For example, prices collected on the web may only represent purchases made online, or they may also cover purchases made in some physical stores (for instance stores with an online presence).

Data integration

Arithmetic or geometric aggregation ?

$$Laspeyres = w_{NewDataSource} I_{NewDataSource} + w_{Field} I_{Field}$$

$$Geometric\ Laspeyres = (I_{NewDataSource})^{w_{NewDataSource}} * (I_{Field})^{w_{Field}}$$

- Laspeyres ≥ Geometric Laspeyres
- The arithmetic variant would be consistent with the higher-level Laspeyres-type aggregation.
- The geometric variant would be consistent with a Jevons index that aggregates price observations across outlets.

Data integration

- A **multi-source price index** that combines sub-indices based on different methodologies.

- Scanner data : Census of transactions (prices and quantities), price index of a high quality

- Web scraped data: large data sets, good coverage, but may still be subject to some ‘bias’ due to the lack of weights

- Prices collected in the field: smaller data sets, the result of multi-stage sampling designs that include non probability sampling methods

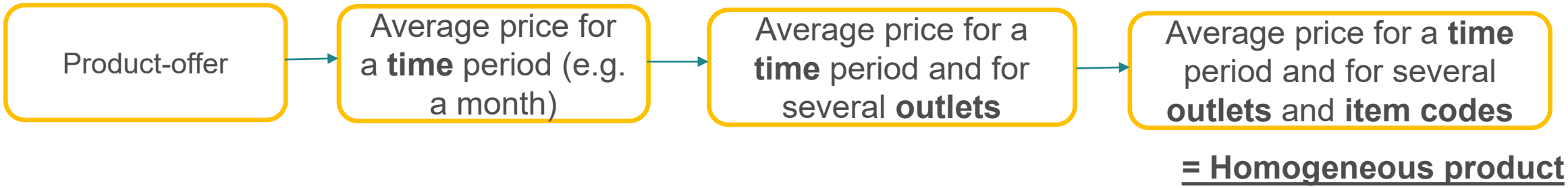
Can the **weight allocation between data sources be optimized**, taking into account the ‘quality’ of each data source?

Data integration

- The replacement of prices collected in the field with prices obtained from other sources will reduce the number of prices available to compile a field price index.

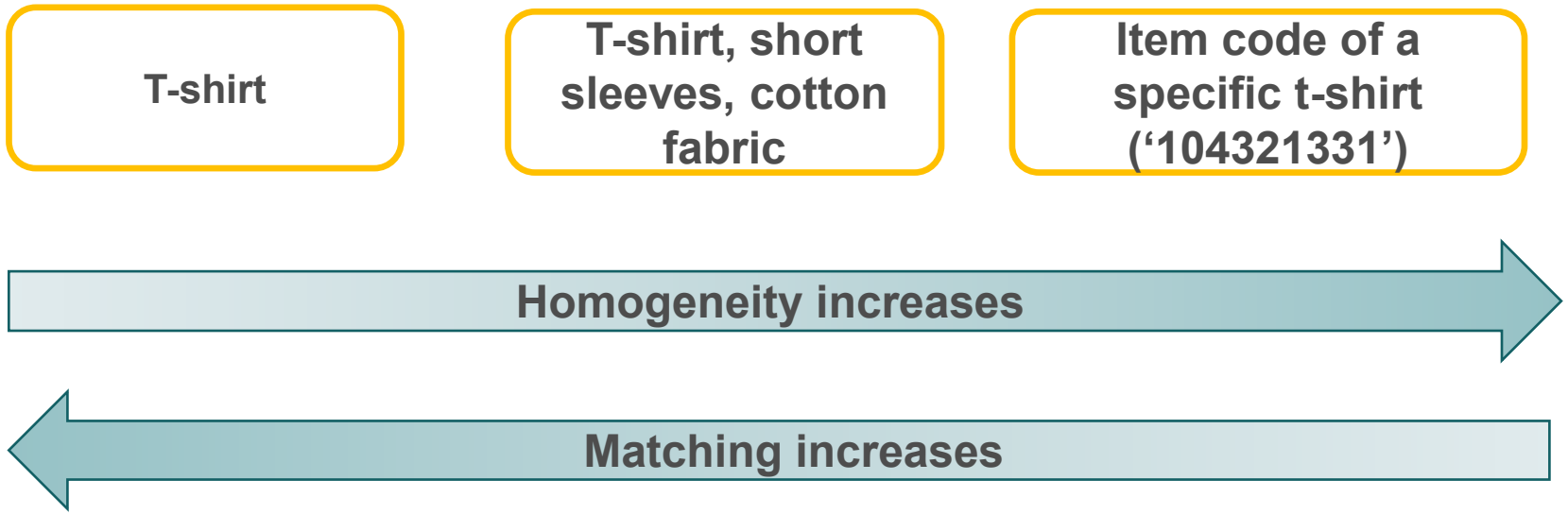
→There can be a need to **redesign the price index for prices collected in the field.**

Homogeneous product



- Three dimensions need to be considered when constructing homogenous products: **time, outlet, product**.
- Assume **no quality difference** between the product-offers that are grouped together.

Homogeneous product



The specification of homogeneous products is a compromise between **homogeneity** (avoid unit value bias) and **increased matching** over time.

Homogeneous product

- This is especially relevant for **clothing**: large and synchronized assortment changes, standardised product features.
- **Metadata** must be available for the construction of homogeneous products: text strings, structured product characteristics, etc.
- The specification of the homogenous product can significantly **impact the end results** !

Web scraping

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source	Web scraping	Laspeyres-type
Product category	Jackets	Bilateral or multilateral index
Homogeneous product	**Jacket of type X and brand Y**	**Arith. or geo. avg. of price quotes**
Product-offer	**Specific jacket at a specific date on a specific website**	**Price observation**

Web scraping

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source	Web scraping	Laspeyres-type
Product category	**Jackets**	**Bilateral or multilateral index**
Homogeneous product	Jacket of type X and brand Y	Arith. or geo. avg. of price quotes (no matching)
Product-offer	Specific jacket at a specific date on a specific website	Price observation

- **Weights:** No weights, or use of proxy weights
- **Bilateral:** chained or fixed base
- **Multilateral**: window length, splicing method

European Commission

Scanner data

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source Respondent	Scanner data Retailer	Laspeyres-type
Product category	Jackets	Multilateral
Homogeneous product	**Jacket of type X and brand Y**	**Unit value price**
Item code	**Specific jacket in a specific month in a specific outlet**	**Unit value price**

Proper **unit values** can be calculated, **conditional on a HP definition**

Scanner data

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source Respondent	Scanner Data Retailer	Laspeyres-type
Product category	**Jackets**	**Multilateral**
Homogeneous product	Jacket of type X and brand Y	Unit value price
Item code	Specific jacket in a specific month in a specific outlet	Unit value price

Scanner data

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source Respondent	**Scanner data Retailer**	**Laspeyres-type**
Product category	Jackets	Multilateral
Homogeneous product	Jacket of type X and brand Y	Unit value price
Item code	Specific jacket in a specific month in a specific outlet	Unit value price

An **elementary aggregate** is the smallest aggregate used in a Laspeyres-type index.

Article 2(13) of Regulation (EU) 2020/1148

Scanner data

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source Respondent	**Scanner data Retailer**	**Laspeyres-type**
Product category	Jackets	Multilateral
Homogeneous product	Jacket of type X and brand Y	Unit value price
Item code	Specific jacket in a specific month in a specific outlet	Unit value price

Alternatives:
- **Geometric Laspeyres**
- **Törnqvist**

Scanner data

Level	Example	Index compilation
ECOICOP subclass	03.2.1.1 Garments for men	Laspeyres-type
Data source Respondent	**Scanner data Retailer**	**Multilateral**
Homogeneous product	Jacket of type X and brand Y	Unit value price
Item code	Specific jacket in a specific month in a specific outlet	Unit value price

The product category is skipped and the multilateral method is applied at a higher level

See: Claude Lamboray (2019). 'Elementary aggregation: A not so elementary story!'. Paper presented at the 16th Ottawa group meeting.

European Commission

Conclusions

- New data sources lead to new aggregation structures.
- Different data sources correspond to different strata.
- With web scraped data and scanner data, one must define the homogeneous product and the different levels for which elementary indices are compiled.
- Different index formulas can be used at different stages of aggregation.
- All these ‘structural’ decisions can have an impact on inflation measurement !

Thank you

European Commission