UNIT VALUES AND AGGREGATION IN SCANNER DATA

TOWARDS A BEST PRACTICE

Jörgen Dalén

Statistical Consultant

CONCEPTUAL ISSUES

- Homogeneity
- Consumption segment and aggregation
- Dynamic vs. Static approach to transactions data
- ----------
- Paper emanates from three papers prepared within a project organised by Eurostat
- Opinions expressed are my own

WHAT IS HOMOGENEITY?

Theoretical answer

- A product is homogeneous if all product offers/ transactions within its specification are equivalent to the consumer
- Three dimensions:
  - Product
  - Outlet
  - Time

System of National Accounts

- "It is obviously impracticable to introduce a degree of disaggregation that would identify each of these types of screw separately and the thought of identifying screws separately from nails and other metal construction materials is already implausible. The problem of non-homogeneity is thus inevitable but may be reduced by considering the level of detail available."

Aim at sufficiently homogeneous

PRODUCT HOMOGENEITY

- A GTIN (EAN) is internally homogeneous
- But several GTINs may be identical or equivalent to consumers
- The relaunch phenomenon can lead to extreme bias if GTIN=product
  - Since an outgoing product with a low price is followed by an equivalent product at a higher price $\Rightarrow$ downward drifting index
  - Homogeneity then has to be defined above GTIN level
- But including different qualities in a UVI aggregate leads to mix bias
- In practice a balance has to be struck between relaunch and mix bias
  - Relaunch bias is downward and sometimes catastophic
  - Mix bias can have any sign and is not necessarily large

OUTLET HOMOGENEITY

- Outlets have different service levels
  - Think of an ice cream in a supermarket freezer vs. one sold at the beach
- Including different outlet types in a UVI leads to mix bias
- Ivancic and Fox (2013): "*The same item sold by different sellers is viewed as homogenous if the price of the item is found to be consistently the same across sellers in the long term.*"
  - Note *long term.* A higher service level = a higher price level across products.
- Retail chains have their own outlet categories
  - Can they be taken as homogeneous with respect to service levels?
- For large countries geography also matters

HOMOGENEITY ACROSS TIME

- Is each subperiod within a month equivalent to the consumer?
  - Is the price level in an outlet the same for each subperiod?
- Higher prices in weekends?
- Higher prices in late evenings?
- So far not common for goods but common for services
- Today homogeneity within month can be assumed for goods?

DATA SUPPLY DETERMINES WHAT IS POSSIBLE!

- Values and quantities can be for a chain of outlets or for single physical outlets.
- Values and quantities can be for a week or a day.
  - In principle, we could imagine also other demarcations in the time dimensions such a whole month or, in the other extreme, an hour or less.
- GTINs are simple identifiers in a number format.
  - The extent and form in which additional information on attributes of the products is available varies from country to country, product to product and from one data provider to another.
- In practice data structures supplied will determine the detailed definition of homogeneity.
  - Do we have to accept what is given to us?

SUGGESTED RECOMMENDATIONS

- **Product-offers (GTINs) shall be considered homogeneous if they have the same use and most consumers are judged to consider them of equal value (choose between them on the basis of price only).**
  - **Furthermore, GTINs with life-cycles that frequently end with price reductions must be combined into larger homogeneous groups with a long duration also where small differences in quality exist.**
- **Product-offers (GTINs), which have different price levels over a longer time period, shall not be considered homogeneous.**

CONSUMPTION SEGMENTS (CS) AND QUALITY ADJUSTMENT

- A CS consists of products intended for broadly the same use.
- Traditionally in the HICP, CS is used for rules on replacements and QA
- When traditional sampling practices are replaced by scanner data, the CS concept could determine **the level of fixity in the aggregation system.**
- CS is a set of heterogeneous product with broadly the same use
- Aggregation to CS needs to account for quality differences
- Churn within CS is big to enormous
  - For many products, GTINs/product-offers are less than 12 months in market

DYNAMIC APPROACH TO TRANSACTIONS DATA

- Current Dutch method based on monthly chained geometric means is the first such method used in actual CPI production
  - Also used in Norway (Belgium, Iceland, Denmark?)
- Research has proposed improved, multilateral methods
- **Makes possible to assign algorithms that are well-defined functions of, in principle, all transactions.**
- Quality adjustments can be built into the index in an automatic, generic way
- Lends itself to harmonisation across countries

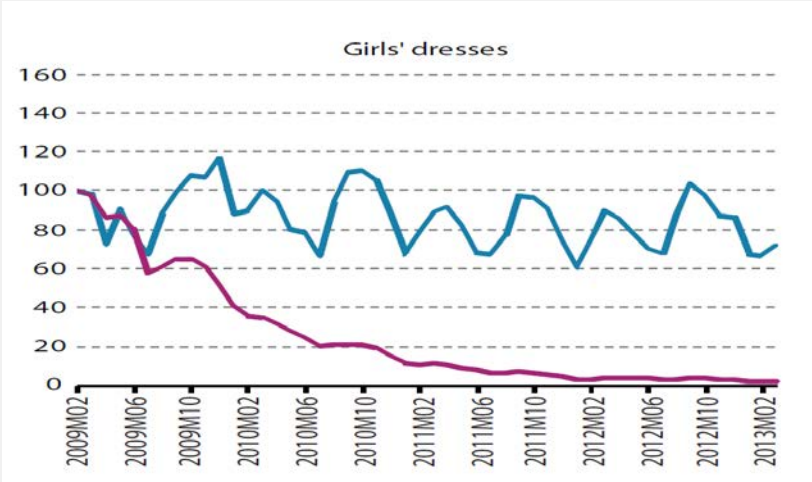
STATIC APPROACH TO TRANSACTIONS DATA

- Used in Switzerland and Sweden
- Tries to extend the Laspeyres fixed base approach down to the lowest (GTIN) level
- Defines GTINs in base period for matches with GTINs in the current period
- Non-matches are dealt with through replacements or deletions
- New GTINs are excluded unless matching old ones
- Where churn is big the method more or less breaks down due to too many replacements in a year
- To a large extent ad hoc methods are needed when GTINs disappear

COMPARISON DYNAMIC- STATIC

The static approaches are similar to traditional methods and as such easier understood	**But traditional methods are not at all easily understood at the detailed level**
Early versions of the dynamic methods were susceptible to chain drift	Present Dutch method and multilateral methods are free from chain drift
The static approach could be seen as applying Laspeyres-type indexes at as low aggregation level as possible	Churn destroys the Laspeyeres approach for most product groups (except maybe for food)
The dynamic approaches could be seen as being more complicated in a mathematical sense	But they are rigorous and well-defined and lend themselves to advanced analyses
Research points in the direction of dynamic (multilateral) methods	

FREE FROM CHAIN DRIFT



Borrowed from Chessa (2016)

- Don’t use non-transitive formulas within year
- Multilaterality – a technique to ensure no chain drift
- Another way is to avoid monthly chaining
- Relaunches can lead to effects similar to chain drift

CONCLUSIONS

- Definitions of homogeneity should be such that relaunches are neutralised, both identical and similar relaunches.
  - Avoiding unit value bias is a second priority by comparison but should nevertheless be addressed to the greatest extent possible.
- Identify the appropriate level for the fixity of the index – consumption segments
- Dynamic methods, which cover the evolving universe of transactions are superior to static methods, which try to assign low level reference units from a historic period.
  - **Why are still some countries choosing static methods?**
- Ensure that methods are guaranteed free of intra-year chain drift.

THANK YOU!

DISCUSSION