Exploring methodologies to integrate new scanner data in the French CPI: making use of multilateral methods

Unece Conference – Geneva

June 2023

Authors: Adrien Montbroussous, Martin Monziols (Insee, France)

Abstract:

Scanner data has been used in production to compute the HICP and the CPI for France since January 2020, for most of French retailers. The current methodology with this data uses a product referential bought from an external provider giving us detailed characteristics for each article. These characteristics allow us to match articles in our data with the COICOP and to create homogeneous groups of articles. Thanks to this information, we can compute a unit price value for each group of articles and each month. The following steps in the methodology are very similar to the process with field-collected data: we select a sample of observations that will be used to compute price evolutions and aggregate them using a geometric Laspeyres formula at the lowest level. And as with field data, replacements are made for unavailable products. This choice is quite specific to France whereas the use of multilateral methods is more widespread in other countries. But recently, new retailers (hard discounters) have started to implement a data flux to provide Insee with their scanner data. The specificity of their data is that most of their articles aren't covered by the product referential, which makes the current methodology hard to apply at first sight.

In this study, the goal is to be able to use these new scanner data in the following years. We will address two questions to do so. First, relying on the other INS experiences, we will test the generalization of multilateral methods on our already received and used scanner data to document on a large scale the behaviours of such methods in the French context. This experience is an opportunity to gain practical and theoretical skills with known data. Second, we will present a strategy to make use of these new scanner data with these methods. Given the raw data of these retailers, the process to be developed goes from classifying the products to integrating the computed indexes in our main process that produces our French CPI.

Contents

I) Context of this study : scanner data in France now and future

After almost 10 years of experimentation, scanner data have been introduced in the French CPI and HICP in January 2020. During these experimentations, first, some retailers were collaborating, then a law has been issued to frame this operation and regulates the transmission of data. Almost all the field of retailers was transmitting data except hard discounters and some retailers in oversea departments.

Recently new providers (hard-discounters) started to send their data as expected by the law. As it will be detailed below, these data are challenging in order to use them in our CPI and HICP given our current methodology.

1) Usual data

In France, scanner data is used in production to compute our CPI since January 2020. Data was until now provided by all the Super and Hypermarket, hard discounter excluded. We are getting data from retailers, thanks to an article of law originally published in 2017, and modified in 2021[1], making mandatory for retailers to provide us data for any day and shop, each day. These are used for prices' statistics and turnover indicators. The data requested is the following:

- EAN (European Article Numbering)
- Outlet id
- Date of the sale
- At least two variables among the 3 following: number of article sold, the whole expenditure and the unit price of the article.
- A label, which can be relatively short and rarely exceeds 25 characters (space included)
- The intern nomenclature code given by the retailer

The law as it is written at the moment implicitly supposes that there is a « referential » that we can use to our purpose of describing and classifying products into a nomenclature. Because, as it is written, the only descriptive information is the label. These data are indeed used in our process with a referential (bought from an external society, CIRCANA previously known as IRI) allowing us to get more information on the data : characteristics of the products and a « family number ». With these data, we are able to classify at a granular level our article (variety, which is even finer than COICOP on 6 positions – a specific level of France) using rules made for each variety to select observations. Lastly, according to the law, we keep 3 years of archives and the data for the current year.

The "SKU" (Stock Keeping Unit) is mentioned in the « guide on Multilateral Methods ». It's a code associated with article that retailers keep to group EAN representing products of the same nature in order to have a better stock and supply process. Unfortunately, we do not possess that kind of information because not specified in the law.

1 https://www.legifrance.gouv.fr/loda/id/LEGITEXT000034540407

2) Scanner data : current methodology

To compute our CPI and HICP indexes with scanner data, we use sampling and a geometric Laspeyres. Thanks to the product referential, we are able to define an « expanded article number » which allow us to keep tract of similar articles, even if they do not keep the same EAN (European Article Number) – it could be interpreted as a SKU code, but for statistical purposes. For instance, if a glue stick changes packaging (with a “Halloween theme” for example), with our extended article number we will be able to keep following the price whereas with only EAN it will be considered as a different product. We are then able to classify these expanded article number into ECOICOP at the variety level. For each group of articles, classification rules of the variety predicted are checked, and if verified the product is linked to the variety. Otherwise it is stashed in the corresponding “poste” (the French specific 6-positions level of COICOP) special category « unclassified ». Each year, operations are made in order to update the basket and the classification rules when it is needed.

There is some sampling because we only follow varieties representing at least 1% of their “poste” (which is an applicative constraint) so that if EANs represent a product not bought enough we do not follow it. In the end, nevertheless, we follow more than 80 millions of EANs. These varieties are updated and modified if necessary each year when we update the basket and the associated weights. These weights are computed with the previous year expenditures as the rest of the CPI basket. Micro-indexes are computed at the outlet x variety scale and then aggregated at the variety level, which is in turn included in the CPI “poste” calculation.

The quality adjustment for replacement is slightly different than the bridged overlap method : since we are able to have an history for the replacement product, we do not need to impute prices of previous periods for the replacement product. The process is automatized for scanner data : we have classification rules that classifies article in homogeneous group and outlay candidates for replacement. Then, among the potential candidates for replacement, we proceed with a sampling with replacement (a product can replace more than one product). Products are replaced if absent for two months in a row. Also, we anticipate some replacements : if a product has both quantity and price declining for two consecutive months, it is replaced the next month.

In each 6 digits COICOP level (“poste”) we have varieties using scanner data and varieties using field collected data. We then aggregate the micro indexes to have an higher level index. For field data, we have micro-indexes computed by geographical areas, for scanner data index we compute the index at the outlet level then aggregate it at the whole France level.

The scope of scanner data used in production is hyper and supermarkets in Metropolitan France, for sales of processed food products, cleaning products and hygiene and beauty products and also, some durable goods. The scanner data expenditure share in the CPI weights is around 10 % of the whole basket. We do not use a larger consumption scope because we do not possess informations in our referential about these products. Hence, we cannot classify these products, we cannot control for their units, etc.

This current methodology was chosen because it was very close to what we do on the field so that our methodology is homogeneous.

3) Information technology infrastructure

These huge datasets (approx. 9 Go per day of received data) is managed with a specific big data infrastructure, divided into:

- A Postgre database containing some information about metadata, referential, nomenclatures, indicators, composition of our scanner data baskets, etc.
- A NoSQL infrastructure for the detailed scanner data, accessible through HUE (Hadoop user experience) on which we can make HiveQL requests. For experimentation and in order to interfere at the least with production processes, we extract subset of the data to explore them with R and dedicated packages. This process is a bit long and laborious since the platform is not designed for such work.

We are on the verge of putting in place a new infrastructure, which will be more flexible and at the state of art, allowing us to foster our experimentation work.

4) New data and data not yet used

a) Hard discounters

We are just in the process of getting data from 2 hard-discounters. Using these data will allow us to have a better coverage of hard-discount and also a better geographic coverage because they are more present in a specific area (north-east). The match rate of these data with the referential is relatively low: for one Saturday of sales, 16,5 % of EANs of one hard discounter are present in our referential, representing approximately 45 % of the data and 40 % of total expenditure. We were not able to compute such statistics with the other retailer since we only have a test file at this stage. Since the data is compliant with what is required by the law, we only have few descriptive information and item labels for each EAN. It has to be noted that the labels seem to be richer than for other retailers, with a lot of products with a 25 character label.

With this situation, we are facing two questions:

- how to classify the product in the COICOP
- how to calculate an index without detailed information about a product (volumes, weight, etc.)

Work will be done on these two subjects, this paper focuses mainly on the later question.

b) Overseas scanner data

As for the hard discounters, we receive and are making new contacts with some retailers to have overseas' scanner data. A large proportion of products are specifically sold in this region of the globe, hence not all the scope of these data is covered by the referential. Then, the same two questions need to be answered: are we able to classify these products with the sole label and is it possible to calculate a good quality price index.

So far we received data for some retailers at La Reunion and are making progress to receive some data from Guadeloupe. It has a relatively low impact on the whole French index but could be source of efficiency and precision for the CPI of these territories[2].

c) Other sectors not yet used

Among the already received scanner data, we restrict our scope to specific products:

- Food
- Hygiene and makeup products
- House cleaning products

2 France does not publish geographic CPI except for the overseas' departments.

- Some durable goods such as pregnancy test, highlighters…

But, for instance, we have data on clothes that we do not use so far. The reason is that this remaining data is not covered by the referential so that products are not classified.

We expect that our work on multilateral methods for the hard discounters will have some positive externalities on existing scanner data not used yet.

II) Theory and strategy of our experimentation

We describe in this section our strategy of experimentation with some reminders regarding multilateral indexes. This work builds on previous experiments with these methods by previous colleagues. They used in their experiment homogeneous product groups using an « expanded article number » created thanks to the referential, on whole milk and foie gras, products that we kept in our study.

In parallel of this experiment, we will start a process to classify hard discounter data in COICOP nomenclature or even granularly (French nomenclature). As for multilateral methods, it will be continuing some previous work done on overseas products.

1) Multilateral methods and milestones of the process

Since our data from hard discounters has only been provided since recently, we decided to start our experiment on the data of the other retailers we already possess. The idea is to calculate indexes based on two assumptions :

- we can classify products at a certain level (COICOP 6 positions for instance)
- we do not possess any detailed information about the product except this classification

In our experiment, we always work at a fixed outlet dimension (at the outlet scale), we will discuss the results according several choices on the product dimension. Since we are dealing in our experiment only with goods, we assume that for the consumer the good has the same utility for each day of the month. We will follow the average price of products by month.

The benefit we expect to have from using a multilateral method is to bypass the classification issues, prevent a basket churn and avoid chain drifts problem caused by bilateral indexes. Instead of comparing only two periods, we will use all the available data within a window of time to compute an index.

Thanks to the multilateral guide produced by Eurostat, we highlighted 4 steps in which we had to analyse several methods/choices, presented below.

a) Individual product specification

In our experiments, we test the following specifications for products and outlets :

- **Products** : as explained earlier, the only identifier that we have in the raw data is the EAN. So the goal here is to document how these multilateral methods behave when we consider the EAN as the identifier of a product or the « expanded article number » which gathers several products that are very similar. These expanded article numbers have been developed in order to capture commercial relaunches and are currently used in our current method. What we want to see is whether we can do without such an « expanded article number » or not. An option not yet explored, and that we will consider if this test is not conclusive, is to build such « expanded article number » based on the available information in the raw data through clustering methods.

- **Outlets** : this dimension has not been fully explored yet. As in the current methodology, we considered the outlets as outlets. We did not aggregate them in any manner, except in the subsection dedicated to contributions below.

At the end, in our data, the way we identify a price and a quantity is at the couple (product x outlet), where product is either EAN or the expanded article number.

b) Multilateral Index

There are several family of multilateral methods:

- Geary Khamis (GK) is a quality adjusted value index. It is an additive method, the index is obtained by solving the following system of equations:

$$I_{GK}^{0,t}=\frac{\sum_{i\in N_t} p_i^t q_i^t / \sum_{i\in N_0} p_i^0 q_i^0}{\sum_{i\in N_t} v_i q_i^t / \sum_{i\in N_0} v_i q_i^0}$$ where v_i inside the window W is $v_i=\sum_{z\in W}\frac{q_i^z}{\sum_{s\in W} q_i^s}\frac{p_i^z}{I_{GK}^{0,z}}$

- Weighted time-product dummy method consisting in an econometric model including dummies for each time period and characteristics. In our context, since we cannot revise our indexes, it is not the more appropriate.
- The last one consists on making transitive these bilateral indexes by averaging across all the possible paths between two dates, inside a time window. It is the GEKS (Gini-Eltetö-Köves-Szulc), which consists of a geometric mean of couple of bilateral indexes:

$$I_{GEKS}^{0,t}=\prod_{l=0}^{T}\left(\frac{I^{0,l}}{I^{t,l}}\right)^{\frac{1}{T+1}}=\prod_{l=0}^{T}\left(I^{0,l}*I^{l,t}\right)^{\frac{1}{T+1}}$$ where T represents the size of the window.

  - GEKS Törnqvist is also called CCD. [3]

In the continuation of this paper, we will focus on GEKS indexes.

- GEKS method is based on bilateral Indexes that are reversible. In our experiments, we consider two of these bilateral indexes :
- **Törnqvist** : Index that is frequent in the literature. It is based on the micro-indexes of products and their relative shares in the expenditures at the two periods of time considered.

$$I_T^{0,t}=\prod_{i\in S}\left(\frac{p_i^t}{p_i^0}\right)^{\frac{s_i^0+s_i^t}{2}}$$

  where the expenditure share of product i in the sample S is $s_i^t=\frac{p_i^t q_i^t}{\sum_{j\in S} p_j^t q_j^t}$. S is the intersection of the basket at time 0 and the basket at time t, i.e. all the products present at both periods.

3 Caves, D. W., L. R. Christensen, and W. E. Diewert. 1982. "Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers." The Economic Journal 92, no. 365: 73-86.

- **Fisher** : a common index with good property. It is the geometric mean of a Laspeyres index and a Paasche index. One for the structure of consumption at the first period of time, the other for the one at the other period of time.

$$I_F^{0,t} = \sqrt{\sum_{i \in S} \frac{p_i^t q_i^0}{p_i^0 q_i^0} \sum_{i \in S} \frac{p_i^t q_i^t}{p_i^0 q_i^t}}$$

c) Time windows & splicing

The GEKS method has some parameters :

- The nature of the window for which we consider the mean of the bilateral indexes
    - Rolling window (each month, the time window is shifted forward by 1 month) with the sub-question of the length of this window : 13 months and 25 months have proved to be useful. The latter has the drawback of needing 25 months before starting to publish indexes but can handle seasonality better.
    - Expansive window (Each month, the time window is extended by 1 month). This method allows to start the production without any background data.

An index between period 0 and a period t can be computed within several windows and hence lead to several results. In order to avoid revising previous indexes, we apply splicing technique to link the index of the latest period with the previous ones. Two choices are possible: using as link the previously published indexes or the recalculated with the new window indexes.

Technically, the splicing is operated via (a) link month(s), that can be

- *Mean splicing* : all overlap periods between the two windows are used in order to link the indexes by computing a geometric average of the pairs of corresponding indexes.
    - Linking with previously (with previous windows) calculated series version

$$I_{pub}^{0,t} = I_{pub}^{0,t-1} * \prod_{k=t-T+1}^{t-1} \left(I_{[t-T,t-1]}^{t-1,k} * I_{[t-T+1,t]}^{k,t} \right)^{\frac{1}{T-1}}$$

- Linking to published series version

$$I_{pub}^{0,t} = I_{pub}^{0,t-1} * \prod_{k=t-T+1}^{t-1} \left(I_{[pub]}^{t-1,k} * I_{[t-T+1,t]}^{k,t} \right)^{\frac{1}{T-1}}$$

- *Half splicing* : period t – ((T+1)/2)+1
    - Linking with previously (with previous windows) calculated series:

- $I_{pub}^{0,t} = I_{pub}^{0,t-1} * I_{[t-T,t-1]}^{t-1,t-(T+\frac{1}{2})} * I_{[t-T+1,t]}^{t-(T+\frac{1}{2})+1,t}$

- Linking to published series:

- $I_{pub}^{0,t} = I_{pub}^{0,t-1} * I_{pub}^{t-1,t-(T+\frac{1}{2})} * I_{[t-T+1,t]}^{t-(T+\frac{1}{2})+1,t}$

In our study we will focus on mean and half splicing, as implemented in the R package IndexNumR[4].

4 https://rdrr.io/cran/IndexNumR/

d) Aggregation structure.

i. There is no way of having a decomposition of the multilateral indexes

We demonstrate here that there is no way to decompose a multilateral index in a sum or product of multilateral indexes. Lets consider a set of product N, which can be decomposed in two subsets N1 and N2.

$$I^{0,t}_{Tornqvist,N}=\prod_{i\in N}\left(\frac{p_i^t}{p_i^0}\right)^{\bar{s}_i} \text{ where } \bar{s}_i=\frac{1}{2}\left(\frac{p_i^0 q_i^0}{\sum_{i\in N} p_i^0 q_i^0}+\frac{p_i^t q_i^t}{\sum_{i\in N} p_i^t q_i^t}\right)$$

Working on $\bar{s}_i$:

$$\bar{s}_i=\frac{1}{2}\sum_{k\in(1,2)} I\{i\in Nk\}\left(\frac{p_i^0 q_i^0\times\sum_{i\in Nk} p_i^0 q_i^0}{\sum_{i\in Nk} p_i^0 q_i^0\times\sum_{i\in N} p_i^0 q_i^0}+\frac{p_i^t q_i^t\times\sum_{i\in Nk} p_i^t q_i^t}{\sum_{i\in Nk} p_i^t q_i^t\times\sum_{i\in N} p_i^t q_i^t}\right)$$

$$\bar{s}_i=\frac{1}{2}\sum_{k\in(1,2)} I\{i\in Nk\}\left(\frac{p_i^0 q_i^0}{\sum_{i\in Nk} p_i^0 q_i^0} p_{Nk}^0+\frac{p_i^t q_i^t}{\sum_{i\in Nk} p_i^t q_i^t} p_{Nk}^t\right)$$

with p_{Nk}^t the share of expenditures of the subset Nk of products in the total set N. As we can see, as long as $p_{Nk}^t\neq p_{Nk}^0$ we can't have something like $\forall i, \bar{s}_i=a I\{i\in N1\}\bar{s}_i^{N1}+b I\{i\in N2\}\bar{s}_i^{N2}$ which would have given this decomposition of the bilateral Törnqvist index :

$$I^{0,t}_{Tornqvist,N}=\prod_{i\in N}\left(\frac{p_i^t}{p_i^0}\right)^{\bar{s}_i}=\prod_{i\in N1}\left(\frac{p_i^t}{p_i^0}\right)^{\bar{s}_i}\prod_{i\in N2}\left(\frac{p_i^t}{p_i^0}\right)^{\bar{s}_i}$$

$$I^{0,t}_{Tornqvist,N}=\prod_{i\in N1}\left(\frac{p_i^t}{p_i^0}\right)^{a\bar{s}_i^{N1}}\prod_{i\in N2}\left(\frac{p_i^t}{p_i^0}\right)^{b\bar{s}_i^{N2}}=\left(\prod_{i\in N1}\left(\frac{p_i^t}{p_i^0}\right)^{\bar{s}_i^{N1}}\right)^a\left(\prod_{i\in N2}\left(\frac{p_i^t}{p_i^0}\right)^{\bar{s}_i^{N2}}\right)^b$$

$$I^{0,t}_{Tornqvist,N}=\left(I^{0,t}_{Tornqvist,N1}\right)^a\left(I^{0,t}_{Tornqvist,N2}\right)^b$$

Hence, a GEKS index cannot be decomposed as the sum or product of GEKS indexes. We may have approximate decomposition (that could be useful for analysis) but we have to choose a level at which we would compute the index that we will publish.

ii. Choosing a level and aggregating these indexes

There are advantages and drawbacks of choosing a high or low level of aggregation. The lower we compute our micro-indexes with multilateral formula and dynamic weight, the harder it is to include new products or outlets during the year, we also apply dynamic weights only at a low level and may not catch well changes of expenditure. However, it introduce stability in the index which makes easier the interpretation and the consistency of it.

In practice, there will be not that much choices for the level at which computing the multilateral index. It will depend on the performance of our classification tool.

As recommended by the Eurostat guide on multilateral methods[5], we would use fixed weights at the subclass level at least. In the French context, in which we publish indexes at a more dis-aggregated level, ECOICOP on 6 positions (postes).

5 **Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Chapter 6, 2022 edition, Eurostat**

2) Test protocol

Given these elements about our current methodology and the multilateral index (GEKS), we aim at testing some elements :

- Is the multilateral index far from the one we publish on a comparable field ?
- Do the results differ when considering the EAN or the « expanded article number » ?
- What does this new method give at the « poste » level ?

In order to proceed we have the following steps :

- Extracting data : given our data infrastructure, we have to construct and extract our data in order to use them with our usual statistical tools rather than coding the index in HiveQL. To limit the time spent in doing so, we choose 3 products. The level of aggregation should allow us to try several methods, we need to extract data at the EAN level. We keep the following information :
  - The unit price (price per unity of volume)
  - The sales (price per article X number of articles sold)
  - The total volume (number of articles sold X volume of each article)
  - The number of articles sold
  - EAN
  - Extended article number
- Choosing products :
  - Milk, because it has a low replacement rate
  - Foie gras, because it has a high replacement rate and high seasonality
  - Lipstick, because it has a high number of EAN by « expanded article number »
  - Then, we generalise at the whole poste to which they belong.
    - Milk : 2 varieties + unclassified.
    - Canned meat : 7 varieties + unclassified.
    - Make-up and care : 6 varieties + unclassified.

III) Results

1) Presence of references across time

Before computing indexes, we looked at the disappearance rate inside each group in order to get some sense of how data behave. This is some useful information to know to understand how indexes will behave on the one hand. On the other hand, this is the kind of side informations that will be useful to index producers in practice.

One measure has a bilateral approach, it is to follow the products sold in January 2020 and check if they are still available the following months.

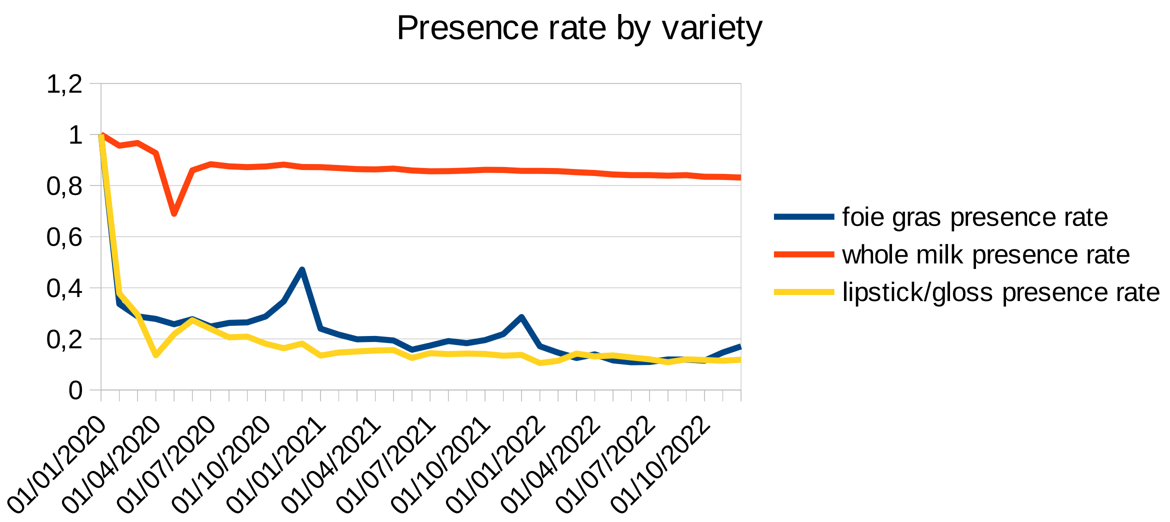

Figure 1: Source: scanner data. Scope: Metropolitan France. Reading note : in January 2021, 13,5 % of the lipstick's EAN sold in January 2020 are still sold in the same outlet

We can see that the product (at the EAN x outlet scale) available in January 2020 disappear rapidly from the market. It makes clear that chain drift is a risk with these methods. When product disappearing is followed with commercial relaunched, EAN change, even if the products are very closed with substantial price rise.

Interestingly, we can see that some products are appearing and disappearing with seasonal patterns. According the variety considered, we can identify a "stock" of products present on the market for many months : approximatively 80-85% for the milk, 10% for the lipstick and foie-gras.

We keep the dates in order to see the clear impact of the Covid-19 crisis, and the lockdown in France.

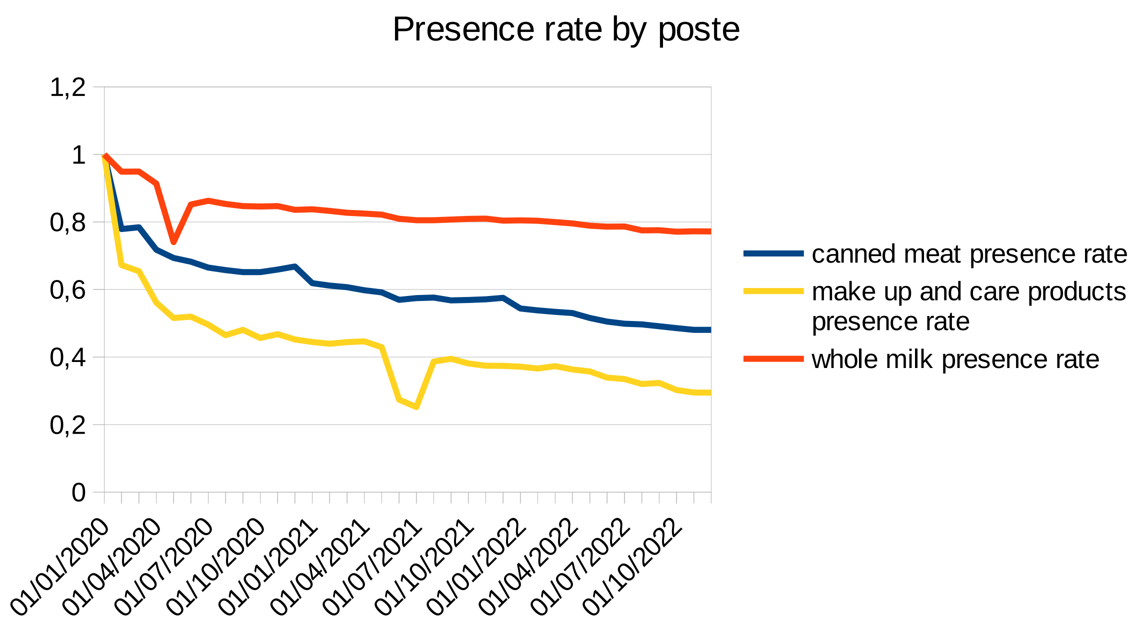

Figure 2: Source: scanner data. Scope: Metropolitan France. Reading note : in January 2021, 61,9 % of the canned meat EAN sold in an outlet January 2020 are still sold in the same outlet

With the same analysis with one level of aggregation, we can see that the trend for canned meat is quite different than the one for foie gras. It is due to the seasonality of the sales that is specific to this variety.

To catch better the matching process that is used in multilateral indexes, we also produced heat maps by comparing the presence of EAN x outlet between each couples of periods within the windows.

a) Milk

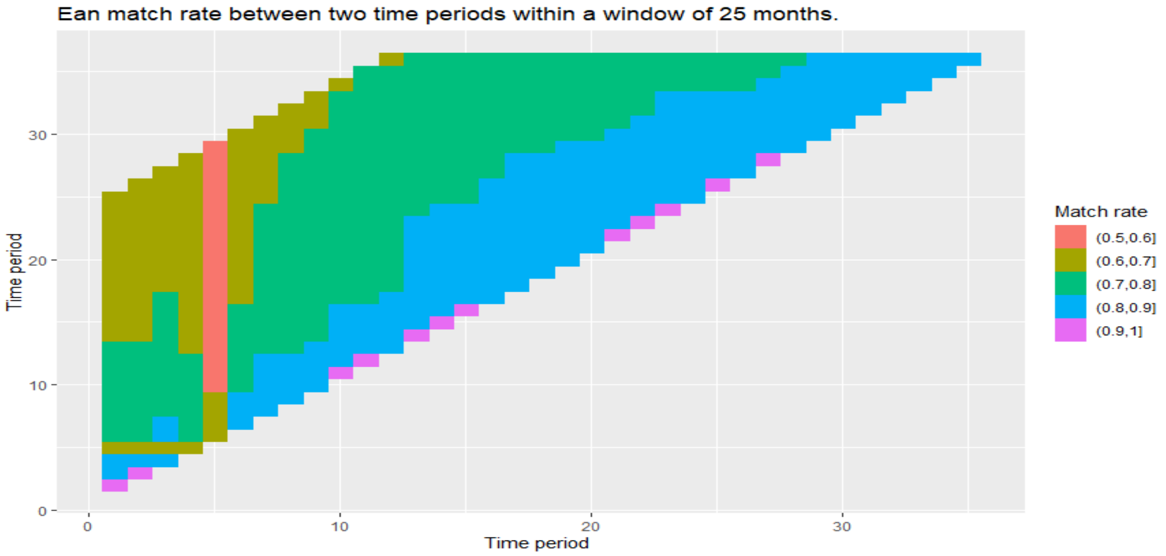

Figure 3: Source: scanner data. Scope: Metropolitan France. Reading note : the quantity represented in the heat map is the EAN match rate computed as (number of EAN X outlet present in both period)/(number of EAN X outlet present in at least one period). Example, 51 % of the EAN X outlet sold in period 1 or 5 are sold in both periods .

The match rate of period 5 (May 2020) is the lowest comparing with other periods. It is more likely related to the lockdown in France following the Covid-19 pandemic.

b) Foie gras

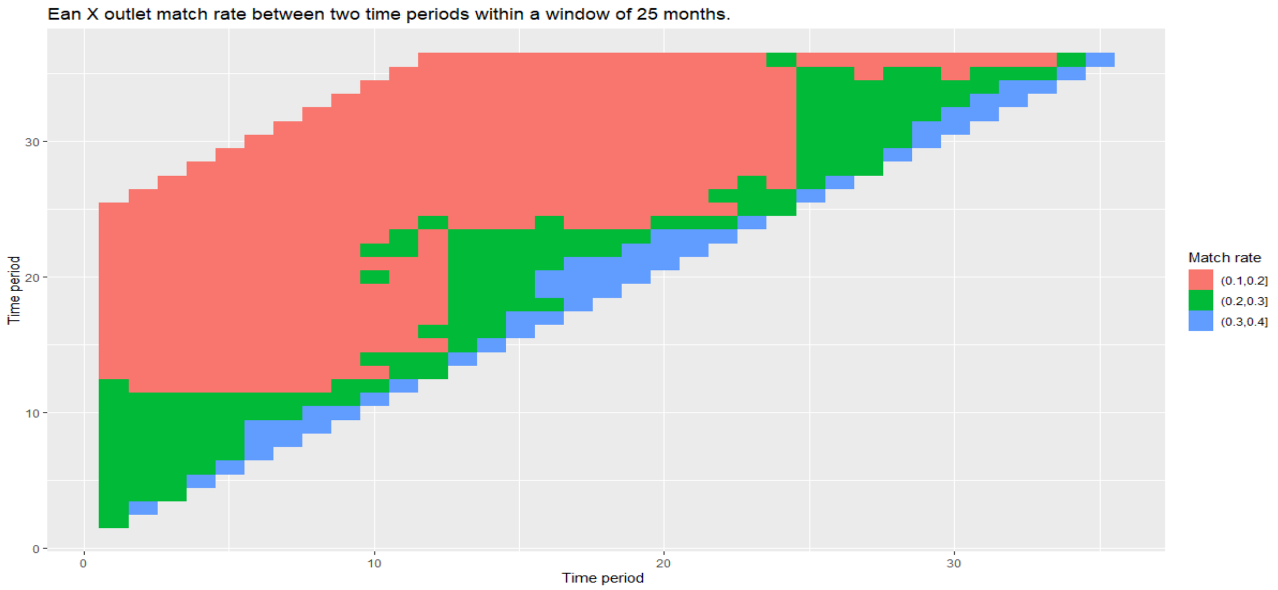

Figure 4: Source: scanner data. Scope: Metropolitan France. Reading note : the quantity represented in the heat map is the EAN match rate computed as (number of EAN X outlet present in both period)/(number of EAN X outlet present in at least one period). Example, between 10 and 20 % of the EAN X outlet sold in period 1 or 5 are sold in both periods .

There is a specificity in the December months (periods 12, 24 and 36) : they have a lower match rate with other months of the year. Indeed, during the winter holidays new foie gras products are introduced into the markets.

c) Lipstick

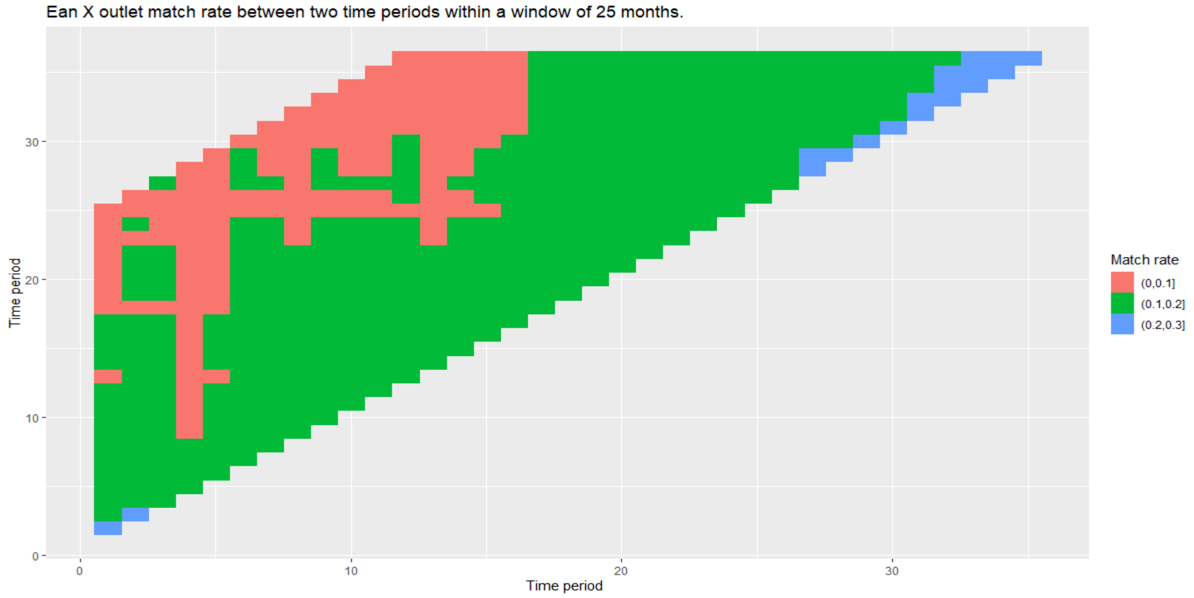

Figure 5: Source: scanner data. Scope: Metropolitan France. Reading note : the quantity represented in the heat map is the EAN x outlet match rate computed as (number of EAN X outlet present in both period)/(number of EAN X outlet present in at least one period). Example: between 20 and 30 % of the EAN X outlet sold in period 1 or 2 are sold in both periods

Period 4 (April 2020) has the lowest match rate with other periods, it is most likely, as for the milk, related to the lockdown.

d) Canned meat

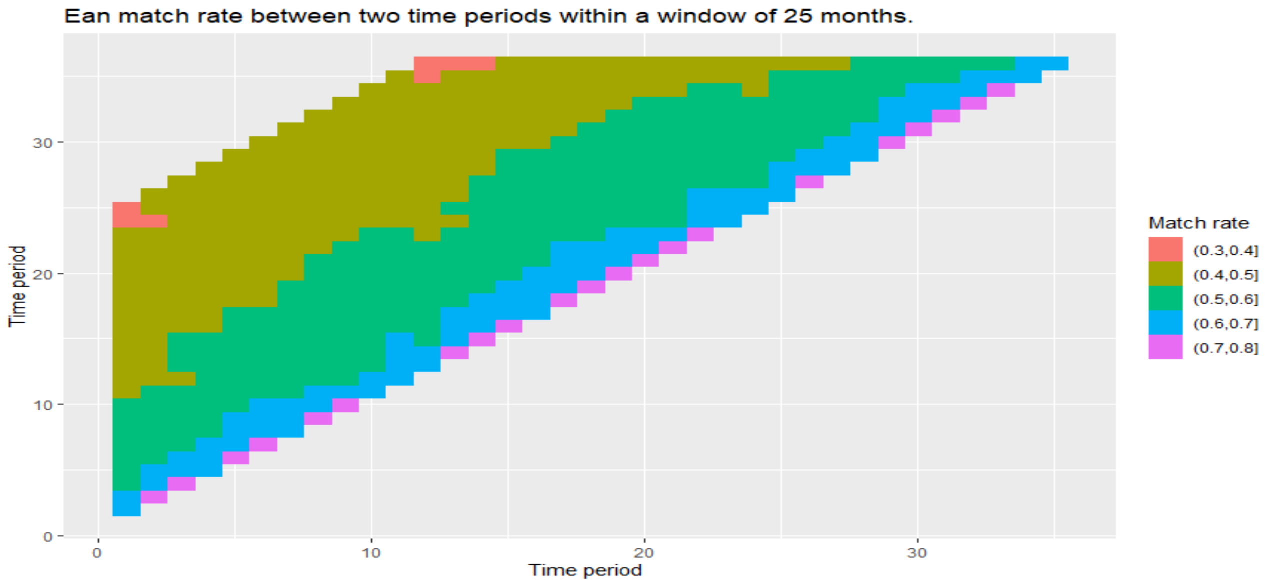

Figure 6: Source: scanner data. Scope: Metropolitan France. Unclassified data are included. Reading note : between 30 % and 40 % of the EAN sold in outlets in period 1 or 25 are present in both periods in the same outlet.

The match rate are higher at the canned meat poste level than for the variety foie gras. An explanation is that most of the varieties does not have the seasonality that foie gras has in the sales.

e) Make up and care products

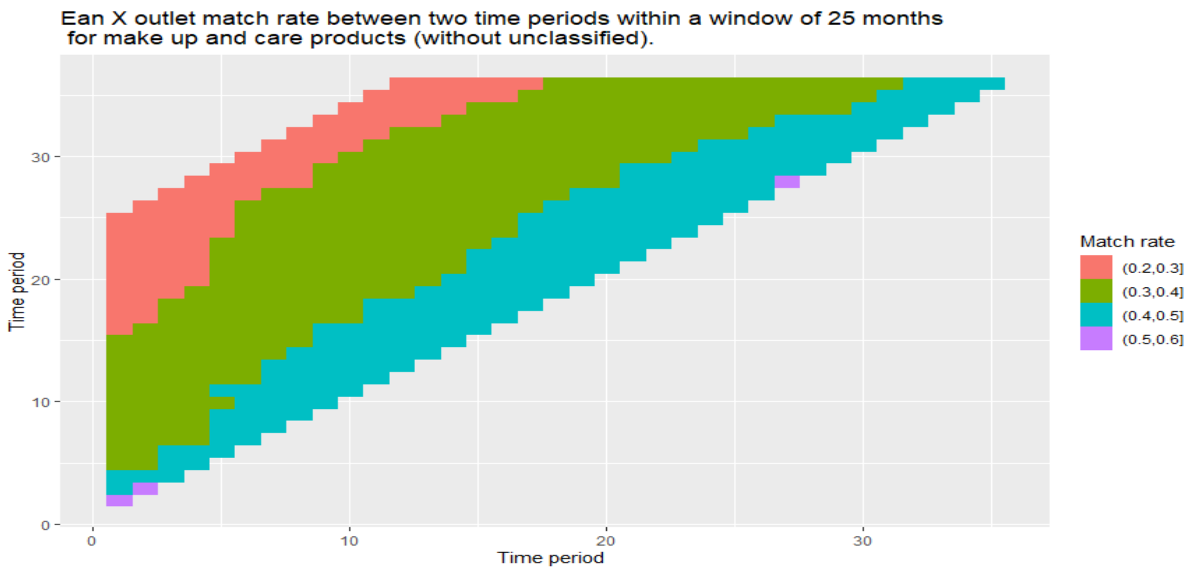

Figure 7: Source: scanner data. Scope: Metropolitan France. Unclassified data are not included. Reading note : between 20 % and 30 % of the EAN sold in outlets in period 1 or 25 are present in both periods in the same outlet.

For the poste make up and care product, the match rate are computed without the unclassified data for reasons of performance and duration of computation.

It seems that there is heterogeneity among products regarding their presence over time. A larger study is needed to have an idea of the scope of possible values of presence over time.

2) Indexes at the « variety » level

For the first comparisons, we only computed the **GEKS Indexes within windows of 25 months**. The idea is to firstly analyse the difference between a multilateral index and our current index and then the differences between using only the EAN and using groups of article (extended article number in our experiment). We used the half splicing method.

a) Whole milk

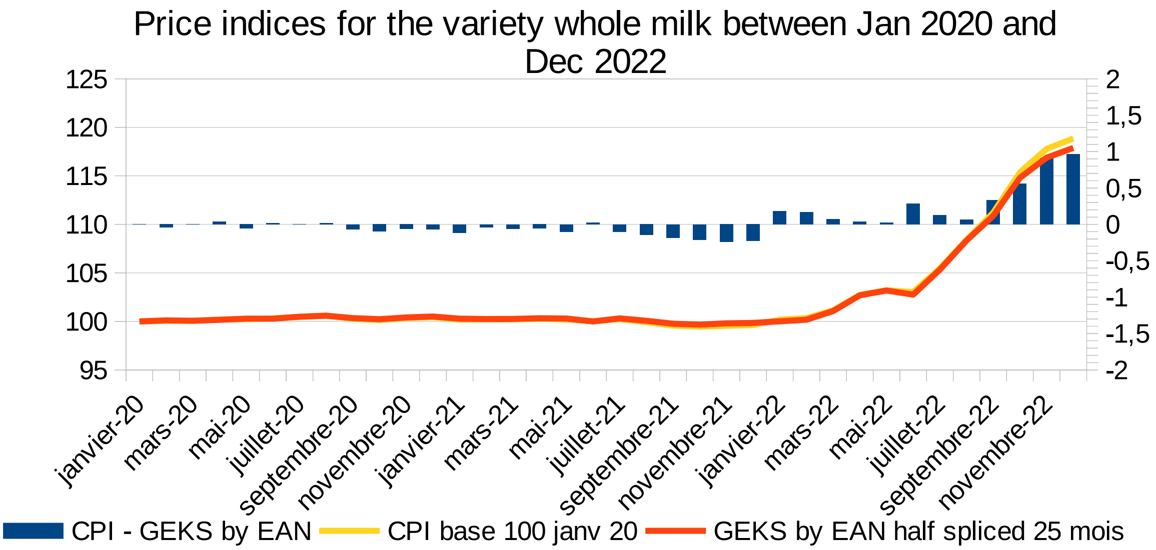

Figure 8: Source : Scanner data Scope : Metropolitan France .Reading note : In December 2022, the price index computed using the GEKS method on article sold grouped by EAN in outlet is 117 it is 1.0 point less than the index computed with a use of « expanded article number ».

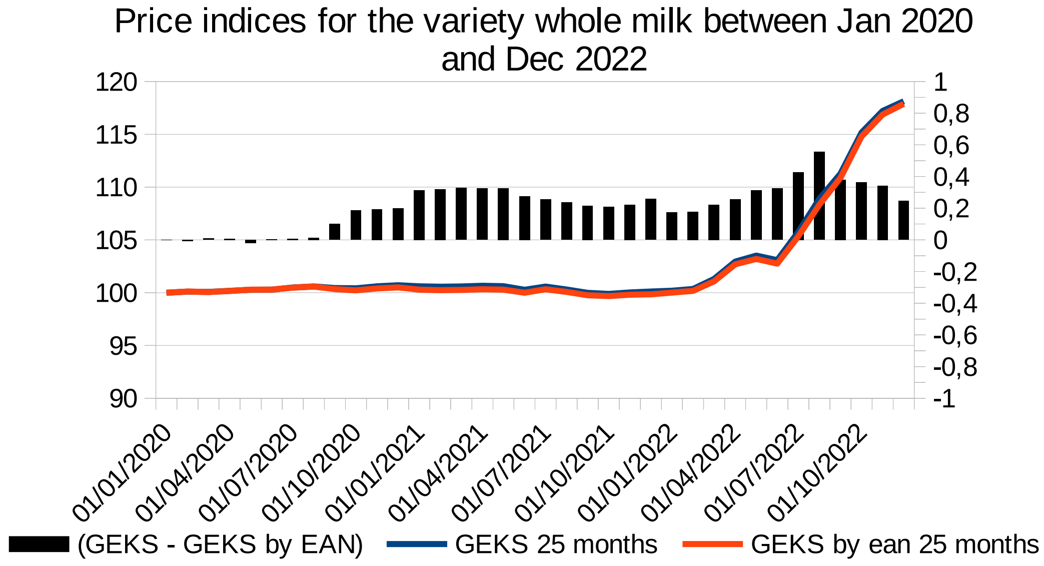

Figure 9: Source : Scanner data Scope : Metropolitan France Reading note : In June 2021, the price index computed using the GEKS method on article sold grouped by EAN is 100.0, it is 0.3 point less than the index computed with a use of « expanded article number ».

Both graphs exhibit very similar price trajectories: between grouping articles by EAN or by « expanded article number » and between a GEKS and the current CPI.

This is working well because whole milk has stable products, few products disappear and the relative shares of the sub-products are relatively stable across time. The price trajectories are the same across all sub-products.

b) Foie gras

During the year 2020, 85% of the products present in our CPI basket in December 2019 were replaced for the variety foie gras.

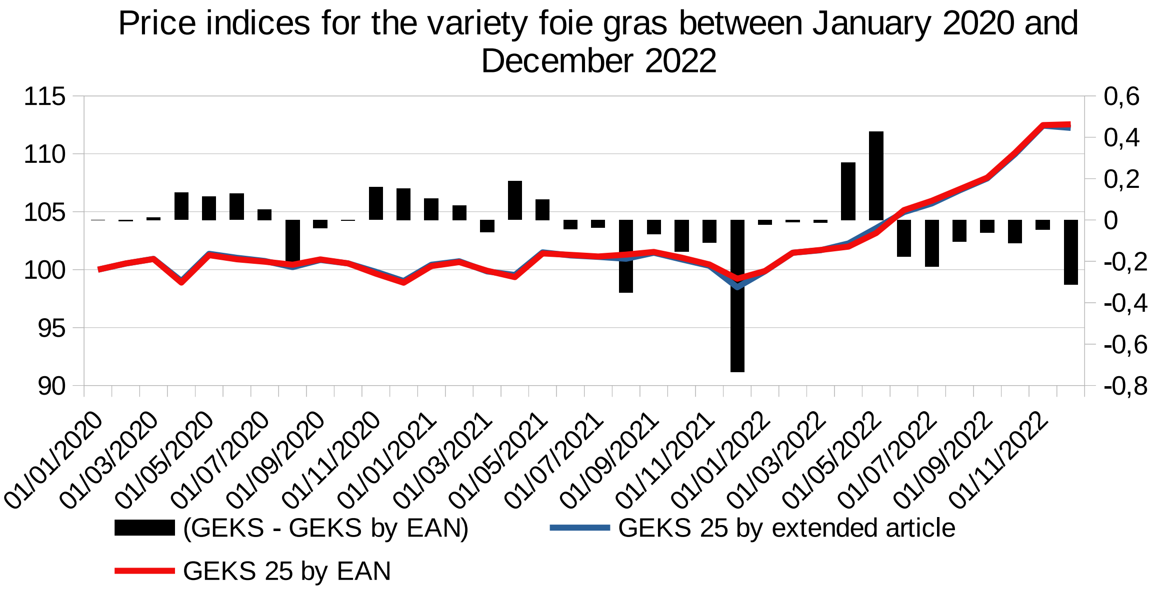

Figure 10: Source : Scanner data Scope : Metropolitan France Reading note : In December 2021, the price index computed using the GEKS method on article sold grouped by « expanded article number » and outlet is 98.48 , it is 0.74 point less than the index computed with EAN .

There is a small difference between grouping the expenditure by EAN or by expanded article number in the case of foie gras. CPI and GEKS index using EAN are relatively comparable, the trend is the same but there is up to 3 points of difference, in a stronger inflation context.

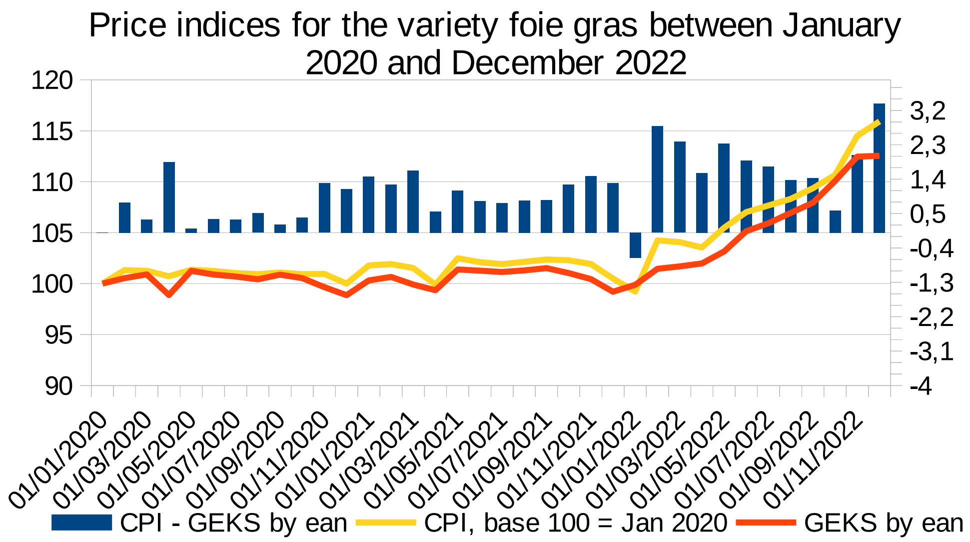

Figure 11: Source : Scanner data and French CPI. Scope : Metropolitan France Reading note : In February 2022, the French CPI (rebased in January 2020) is 104.25 it is 2.78 points more than a GEKS index computed with a use of EAN.

These small differences are even smaller if considered with year-to-year inflation.

c) Lipstick / gloss :

For lipstick and gloss, each expanded article number gathers a high number of EAN: 775 « expanded article » representing 5564 EAN.

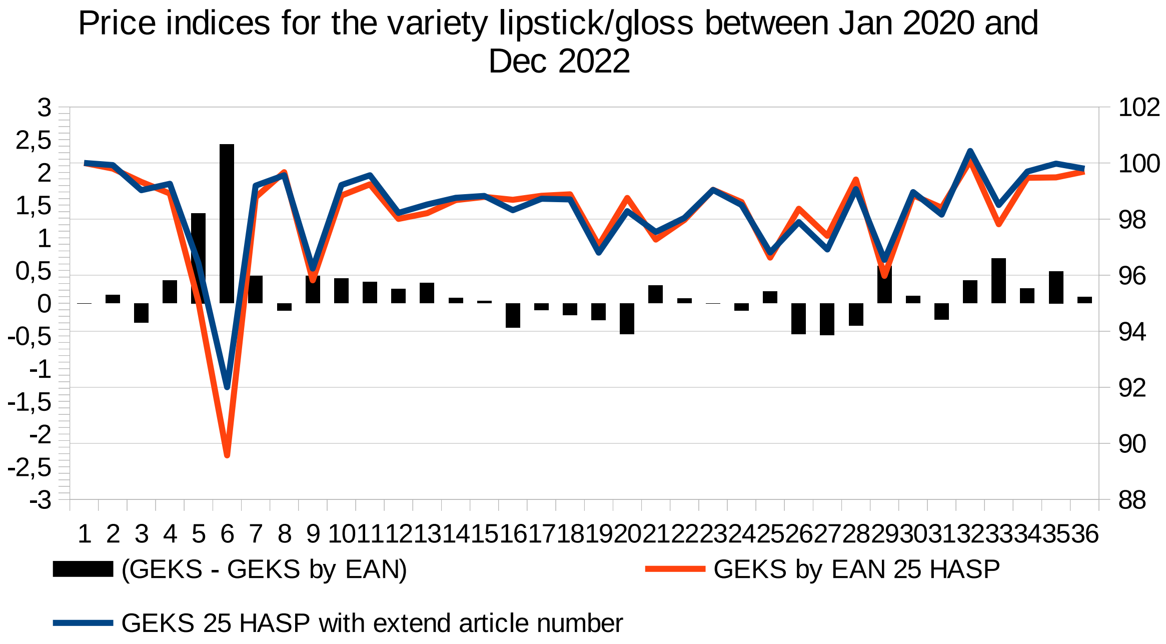

Figure 12: Source : Scanner data. Scope : Metropolitan France. Reading note : In June 2020, the price index computed using the GEKS method on article sold grouped by EAN with a window of size 25 and the half splicing method is 89.6, is it 2.3 point less than the index computed with a use of « expanded article number ».

This first comparison gives similar results with a bit more volatility with the index constructed at the EAN level.

We analysed at the expanded article level the price dynamic and expenditure share to understand better the dynamic, graphics are available in appendix, figures 29 and 30.

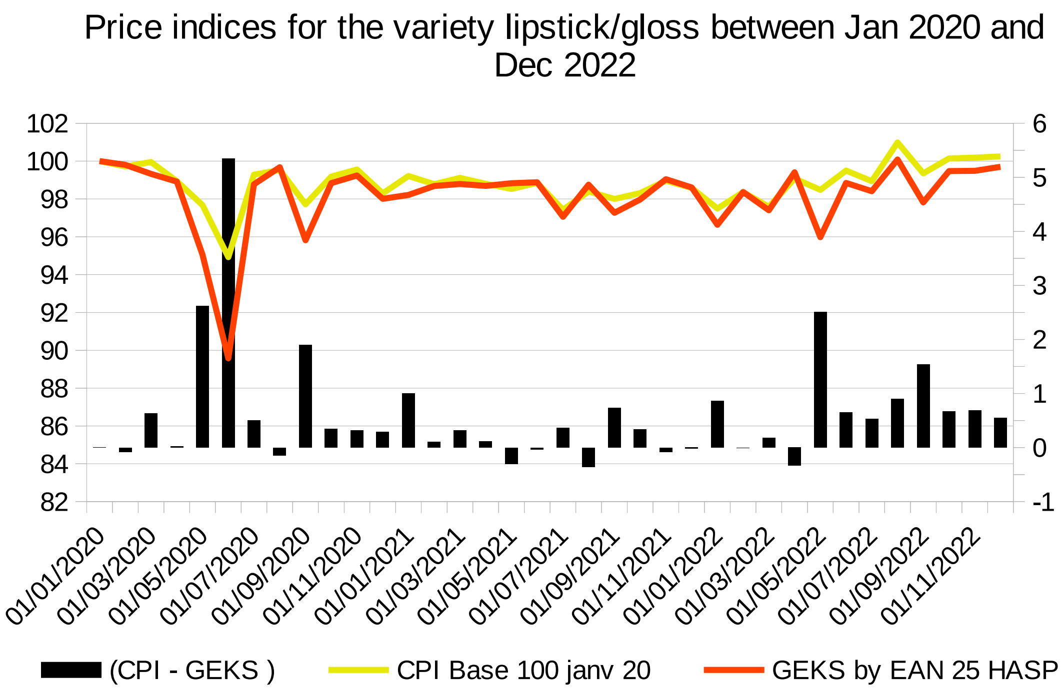

Figure 13: Source : Scanner data and French CPI. Scope : Metropolitan France. Reading note : In February 2022, the French CPI (rebased in January 2020) is 104.25 it is 2.78 points more than a GEKS index computed with a use of EAN x outlet.

The index here again are giving globally the same trends but larger differences than for the 2 other examples. The GEKS is more subject to volatility: each drop is a bit stronger.

The largest difference is July 2020, where some COVID-19 consequences are probably at stake.

3) Indexes at the « poste » level

In our current methodology, in "poste level", there are varieties using scanner data and varieties using field collected data. They are the aggregated together using an arithmetic Laspeyres.

a) Whole Milk

This table presents the weight distribution among all the varieties regarding whole milk – from scanner data and field collected data, the one from scanner data are prefixed by "DC".

YEAR	Label	WEIGHT
2020	WHOLE MILK PASTEURISED	9
2020	Whole milk UHT	17
2020	DC_Whole Milk	60
2020	DC_Fresh pasteurised whole milk	14

The scanner data weight 74 % in 2020 in our poste index. In our raw data, we have 2 varieties that are included in the index compilation and some unclassified data, not used. The data size of 3 years, aggregated by EAN X Outlet X Month represents approximatively $4,3*10^6$ lines.

Expenditure share of varieties inside the whole milk poste between Jan 2020 and Dec 2022.

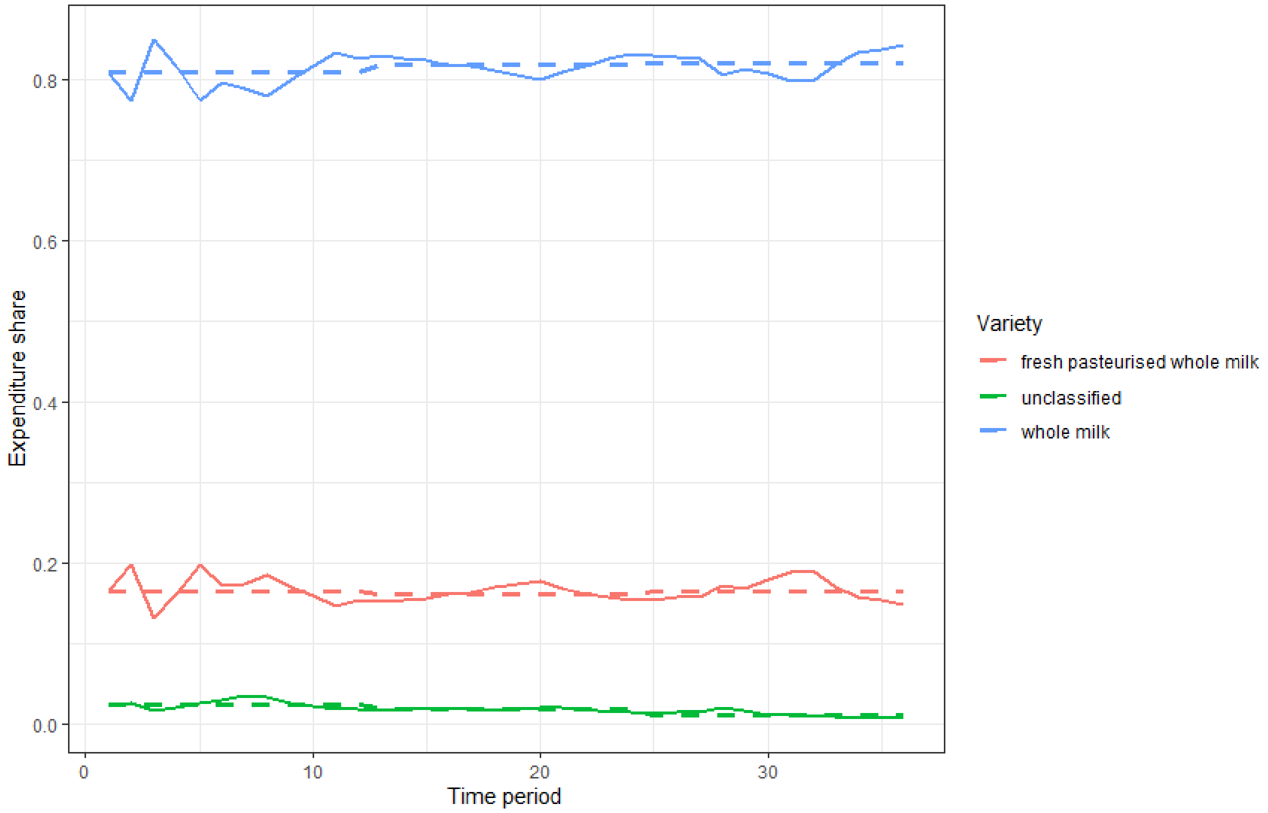

Figure 14 Source : Scanner data. Scope : Metropolitan France. The dotted lines represent annual expenditure shares and the continuous one monthly shares.

The variety whole milk represent the large majority of the scanner data varieties in the poste whole milk. The unclassified products, are almost negligible.

The thing with these unclassified data is that we won't be able with our classifying tool to have such non stable and excluded data. We will maybe have some unclassified observations because our tool won't be able to classify them with enough confidence but with no guaranty that it will be same kind of products.

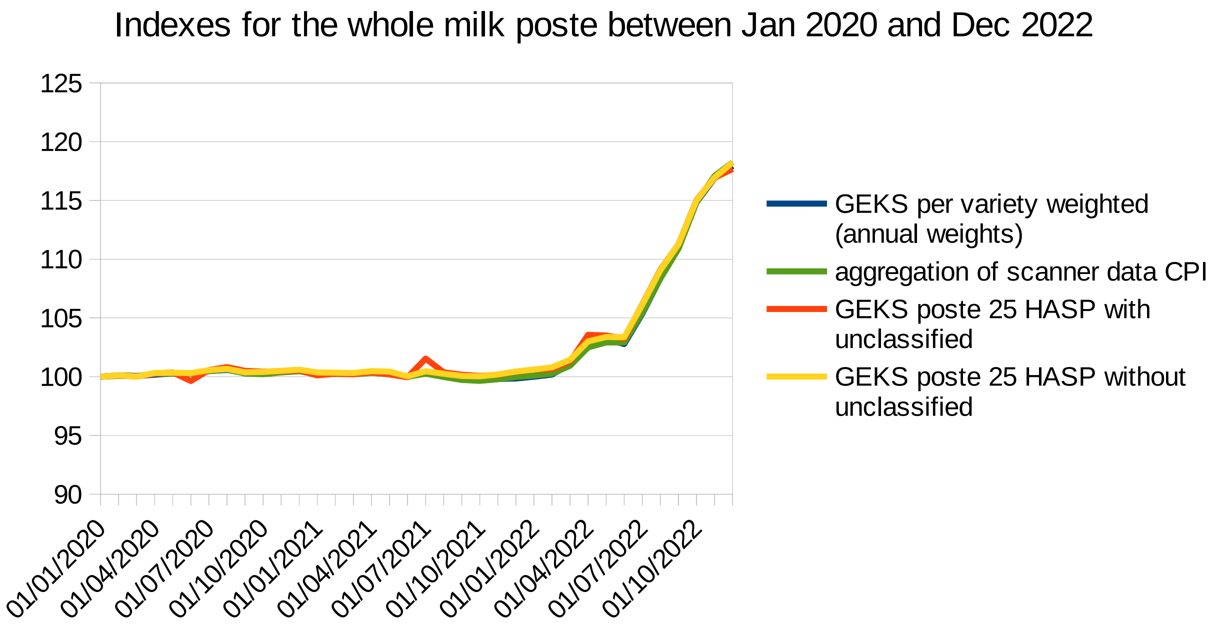

Figure 15: Scanner data Scope : Metropolitan France. Reading note : In July 2021, the GEKS index computed with a window of 25 months grouping by EAN x outlet, half splicing method and including the unclassified data is 101.5. The GEKS indexes are computed with a Törnqvist index formula, the splicing method is mean for the window size 13 and half for the window size 25.

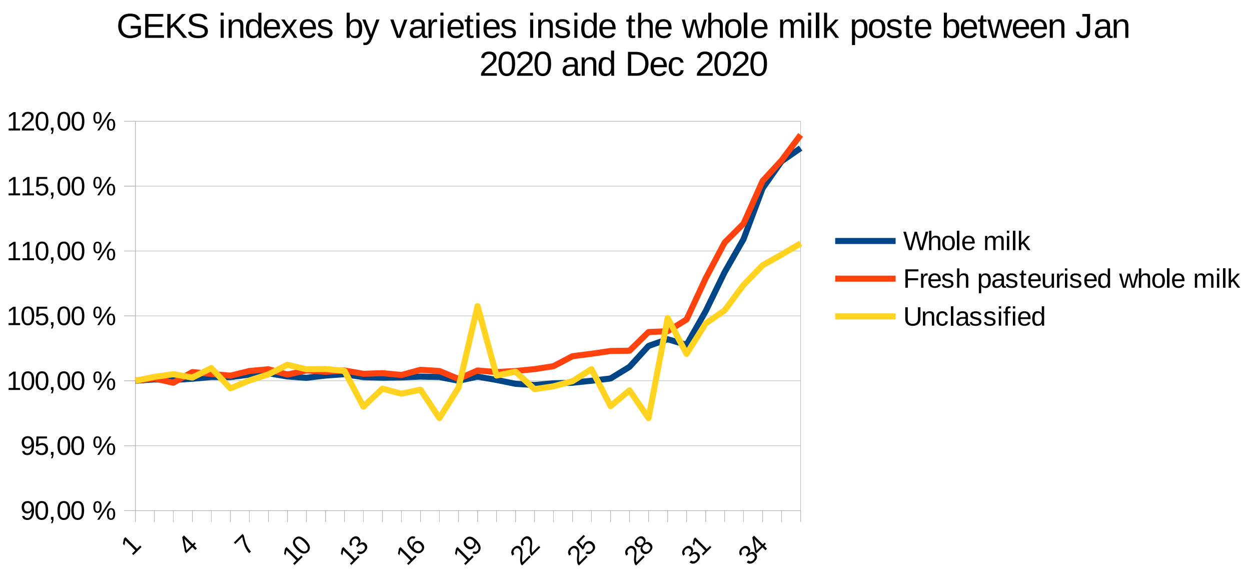

Figure 16: Source: Scanner data. Scope : Metropolitan France. Reading note : In period 28 (April 2022), for the variety fresh pasteurised whole milk the GEKS index computed with a window of 25 months grouping by EAN x outlet and half splicing method is 103.75.

The unclassified products have a more erratic price variation, but they weight very lightly in this poste. It explains the fact that the several GEKS indexes lead to very close results at the whole milk poste level.

b) Canned meat

Scanner data weights 65% in 2020 in our poste index, it represents 7 varieties that are included in our index and unclassified data, which weight more in this “poste” than for whole milk.

The data size of 3 years aggregated by EAN X Outlet X Month is approximatively 22,5*10^6 lines

In order to understand what weights more in the indexes variation, we firstly looked at the monthly and annual expenditure shares of each varieties within the poste canned meat.

Year	Label	WEIGHT
2020	Canned charcuterie	35
2020	DC_Canned rillettes	4
2020	DC_Canned duck confit	20
2020	DC_Canned country style pâté	19
2020	DC_Canned liver pâté	4
2020	DC_Canned poultry pâté	3
2020	DC_Canned full foie gras	9
2020	DC_Canned bloc of foie gras	6

Expenditure share of varieties inside the canned meat poste between Jan 2020 and Dec 2022.

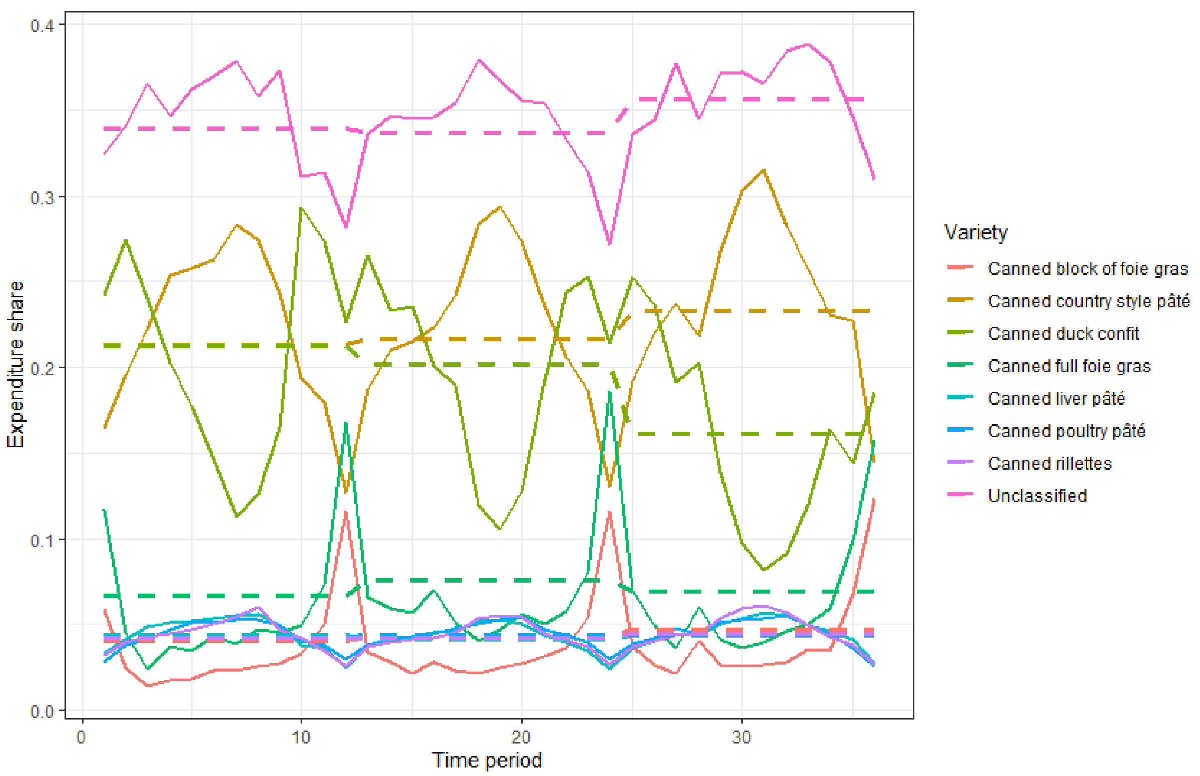

Figure 17 Source: Scanner data. The dotted lines represent annual expenditure shares and the continuous one monthly shares.

We can see the seasonality in the sale of some varieties :

- Foie gras are more sold during the end of the years (December principally).
- Country style pâté & unclassified are less sold in December.

Unclassified data has the most important weight in all periods (approx 35% annually), it is really different than for milk.

We wanted to investigate more these unclassified data, to do so we used the nomenclature we have from Circana which provide us with the referential of products.

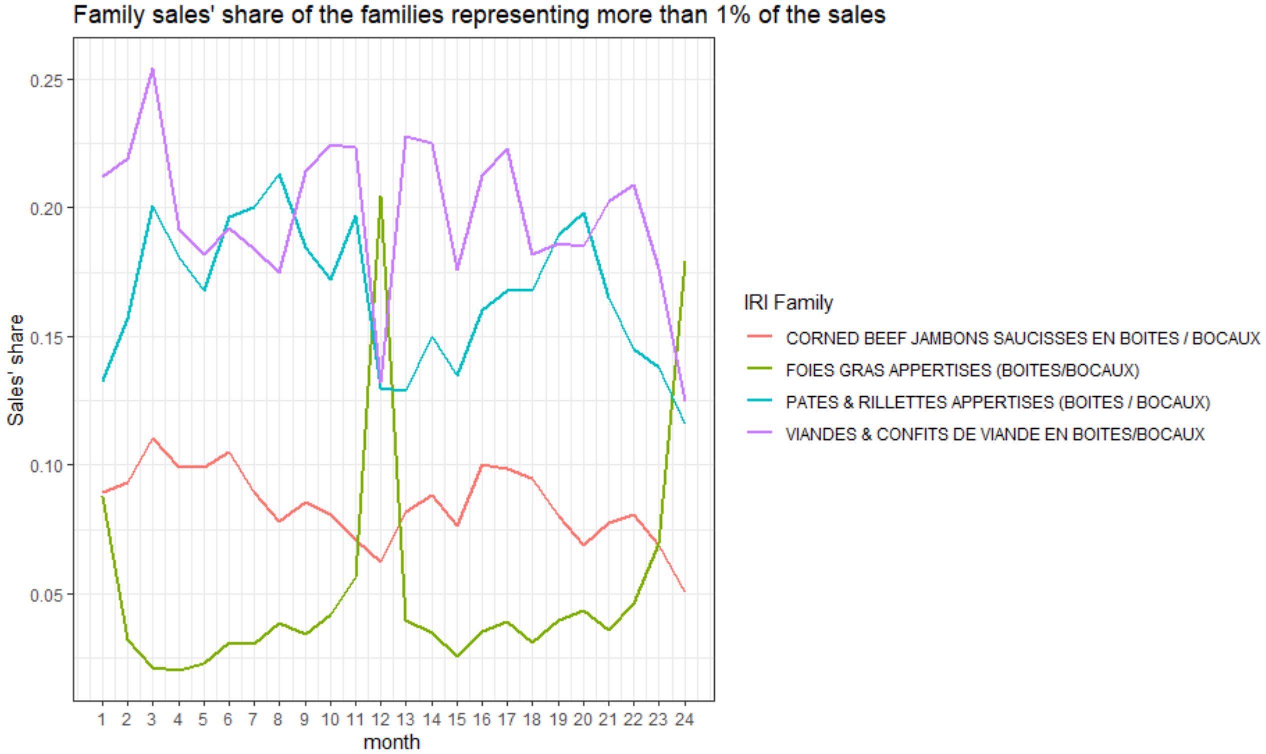

Figure 18: Source: scanner data in 2020 and 2021. The dotted lines represent annual expenditure shares and the continuous one monthly shares. Reading note: in January 2020, the Circana family tinned foie gras represented 8.6% of the expenditure of unclassified data in the poste canned meat.

The EAN represented are part of 4 different "Ciracana families" (a specific nomenclature). Among these families, one could be linked to a field collected variety: "Corned beef and ham". It weights less than 10% of the products in most periods, including this data in our computation could induce "double counts" with the field variety and lead to an overestimation of the weight of the variety canned charcuterie.

GEKS indexes by Circana family in 2020 for unclassified data at the canned meat poste level

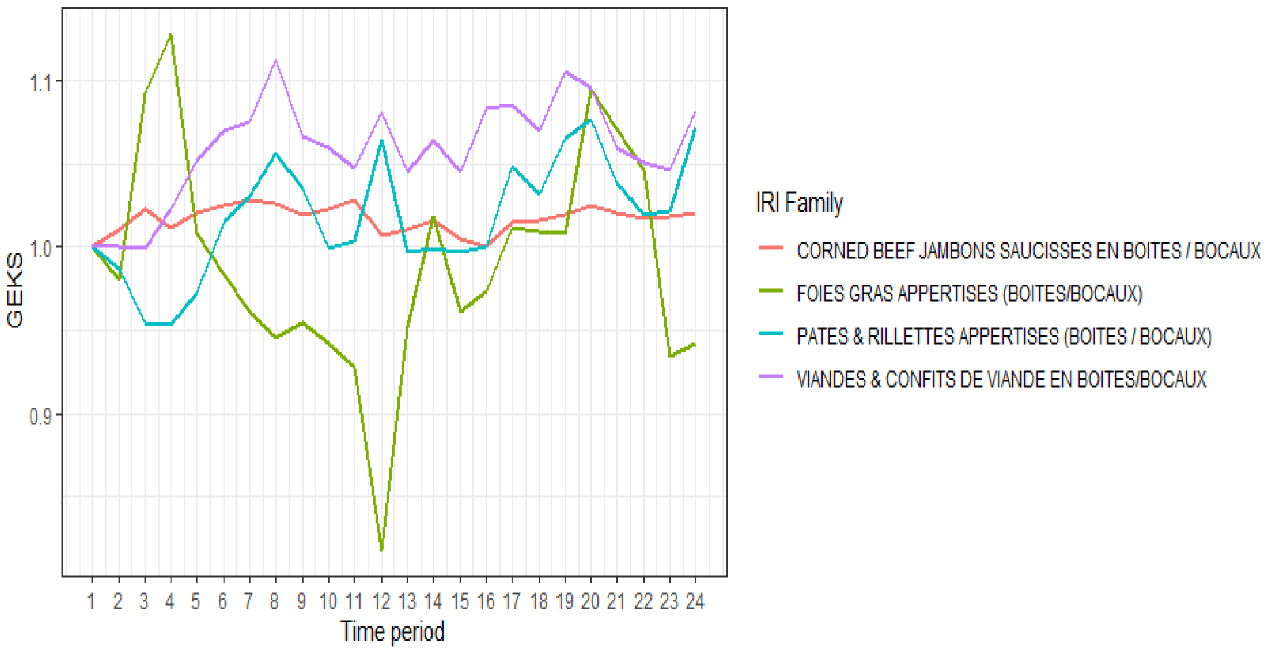

Figure 19: Source: scanner data. Scope: Metropolitan France. Reading note: in December 2020, GEKS index for the Circana (previously IRI) family "canned pâtés and tinned rillettes" grouping by EAN x outlet with a window of 25 month and half splicing was 105.2.

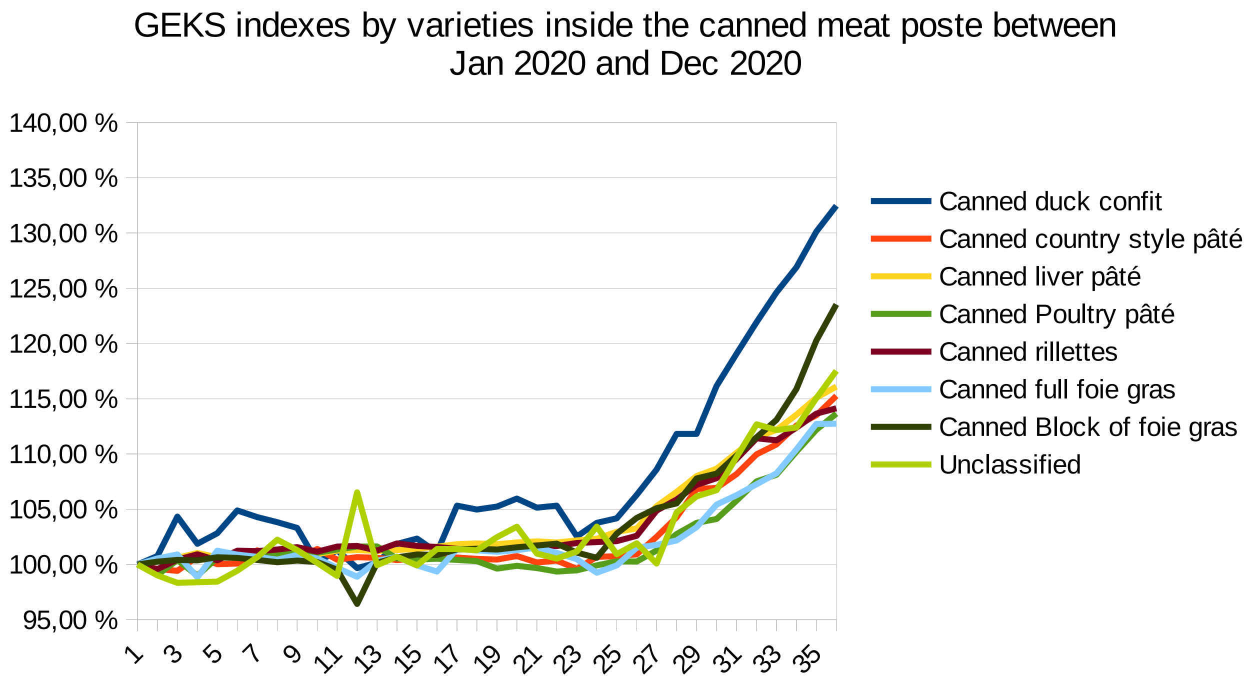

Figure 20: Source: scanner data. Scope: Metropolitan France. Reading note: in December 2020, GEKS index for the unclassified data among the poste canned meat grouping by EAN x outlet with a window of 25 month and half splicing was 106.5.

There is an increase of the index for unclassified data in December 2020. Thanks to the Figure 18, we can see that it is most likely due to unclassified pate, rillettes and confits.

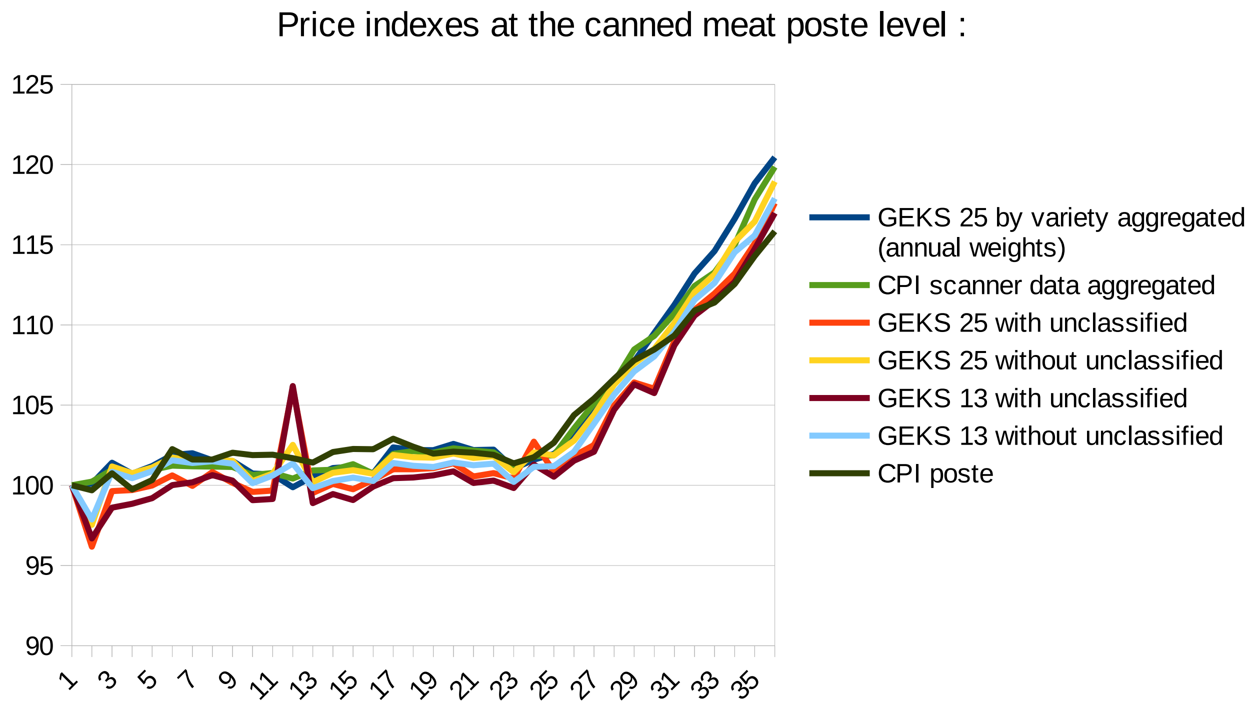

Figure 21: Source: scanner data and French CPI. Scope: Metropolitan France. Reading note: in period 12 (December 2020), GEKS index for the poste canned meat including the unclassified data grouping by EAN x outlet with a window of 13 month and mean splicing was 106.2. The GEKS indexes are computed with a Törnqvist index formula, the splicing method is mean for the window size 13 and half for the window size 25.

The increase in the index including unclassified data in December 2020 is still present. All the indexes are relatively close and have the same trend, except for this month.

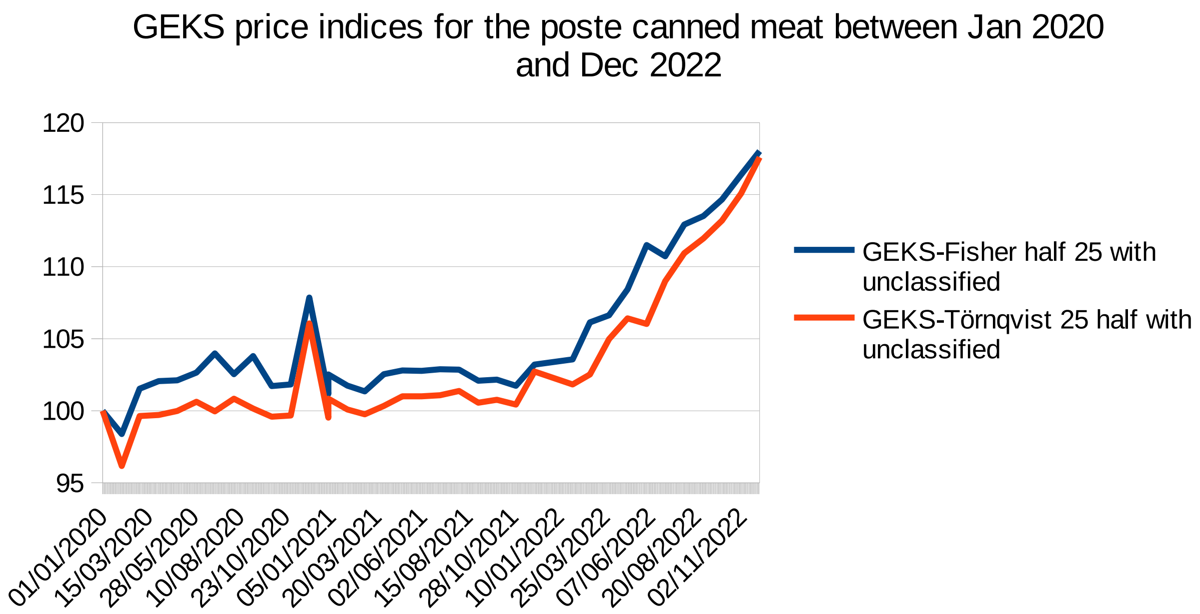

Figure 22 Source: scanner data. Scope: Metropolitan France. Reading note: in December 2020, GEKS index for the poste canned meat including the unclassified data grouping by EAN x outlet with the Fisher Index is 106.1. The GEKS indexes are computed with a window size of 25 and half splicing method.

We also compared the results between a GEKS-Törnqvist and a GEKS-Fisher. It gives relatively similar results, Fisher index seems to lead to higher values.

c) Make-up and care products

Year	Label	Weight
2020	Lipstick	2
2020	Face powder	5
2020	Nail polish	4
2020	Sun products	6
2020	Cleansing milk	10
2020	Care cream	16
2020	MASCARA	5
2020	Depilatory products	5
2020	Body moisturising milk	7
2020	DC_Face women care cream	17
2020	DC_Face cleanser	6
2020	DC_Body care cream/milk	5
2020	DC_Mascara	5
2020	DC_Face powder	4
2020	DC_Lipstick and gloss	3

Scanner data weight 40% in 2020 in our “Make-up and care products“ poste index, it is composed of 6 varieties contributing to the publish CPI and also some unclassified data. The data size of 3 years aggregated by EAN X Outlet X Month is approximatively 148,1*10[6] lines.

<u>Expenditure share by variety for the poste make up and care product between Jan 2020 and Dec 2022</u>

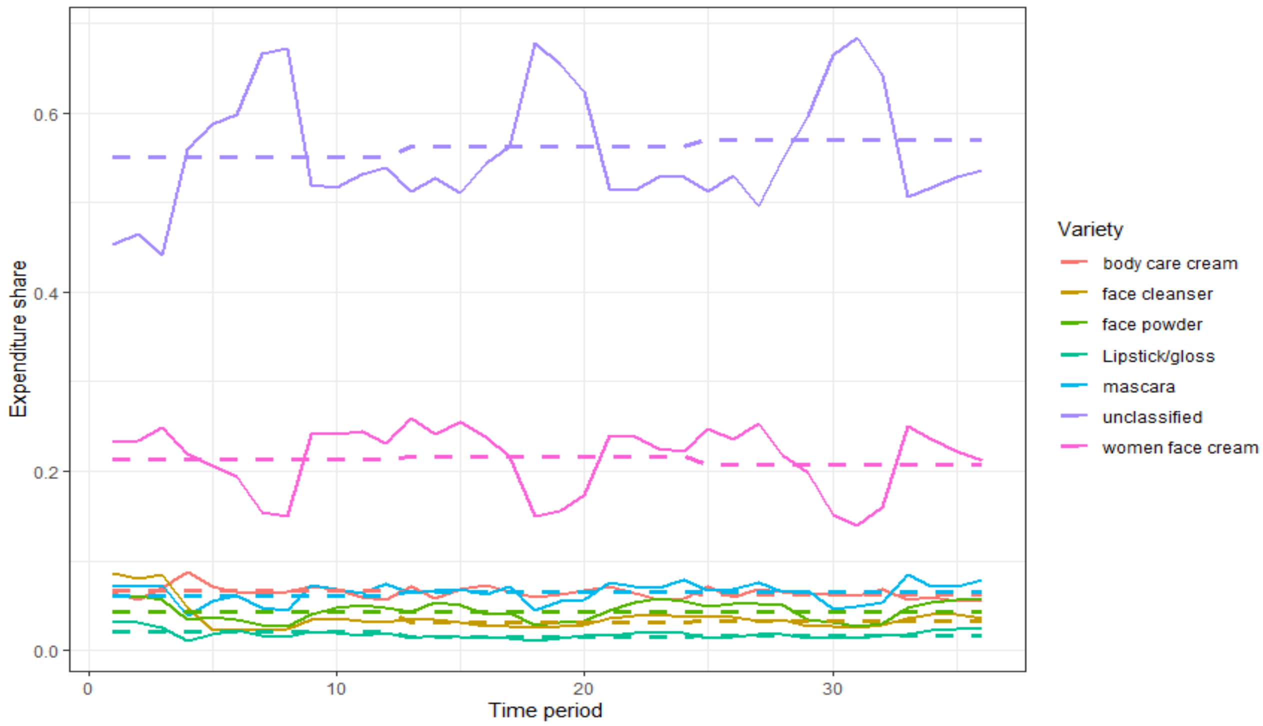

Figure 23 Source: scanner data. Scope: Metropolitan France. Reading note: in December 2020, the expenditure share of unclassified data among the poste make-up and care products is 53.1%

Here also, unclassified data weight a lot, with a strong seasonality. Given that, we can anticipate that at the poste level, with this unclassified data, we could have something quite different from our published index: it weights a lot and has some seasonal pattern.

The unclassified has a high weight for several reasons. First, it is not always easy to make homogeneous class of products. Second, there is an applicative constraint which is that a variety has to be at least 1% of a “poste” so that homogeneous class of products have to gather enough expenditure shares. Third, with time available, the most promising unclassified are prioritise. Hence, some are not studied.

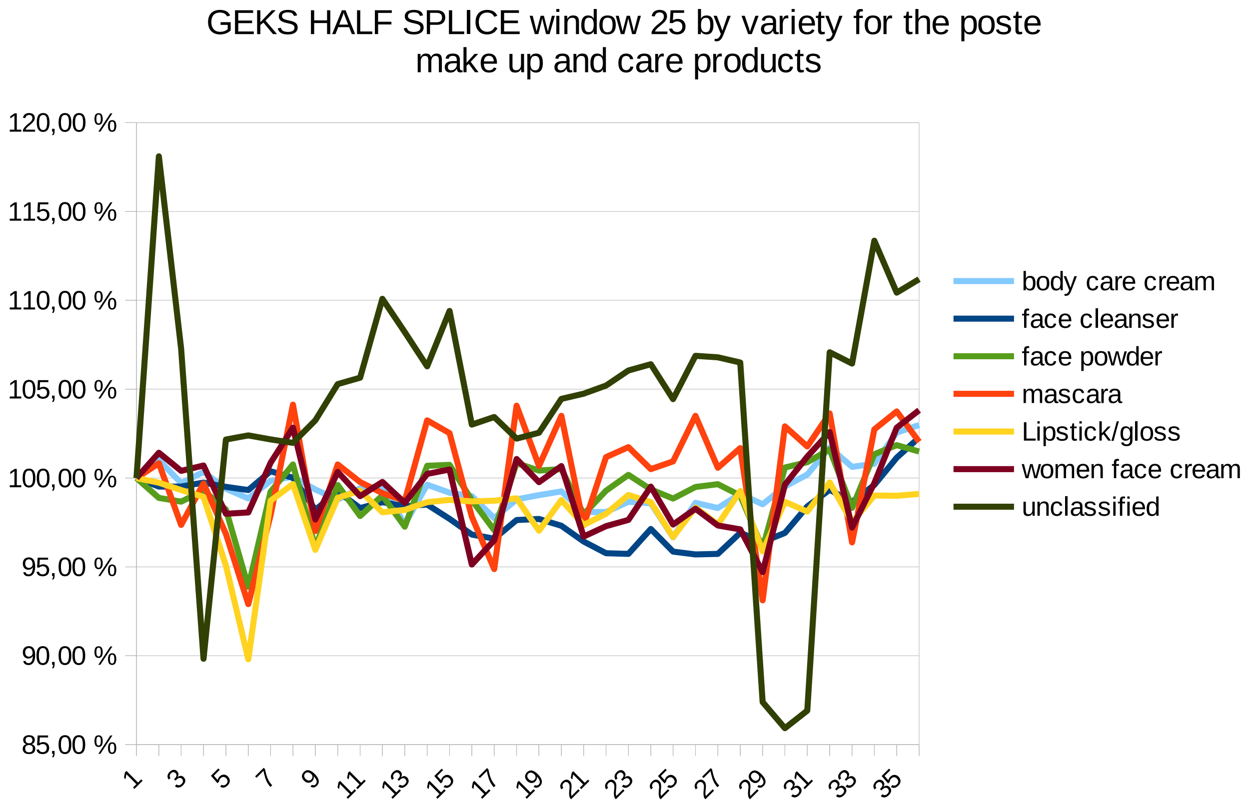

Figure 24: Figure 24 Source: scanner data. Scope: Metropolitan France. Reading note: in December 2020, the GEKS index using a window size of 25, half splicing and EAN x outlet level for unclassified data among the poste make-up and care products is 118.1.

Here we have the multilateral indexes at the “variety” level. The unclassified exhibits some weird behaviour. There are probably some micro-trajectories very steep that have some macro-impact. This case is of interest: we have to develop tools to elucidate that kind of observations: either to understand what it is going on or to cancel these observations if not reliable.

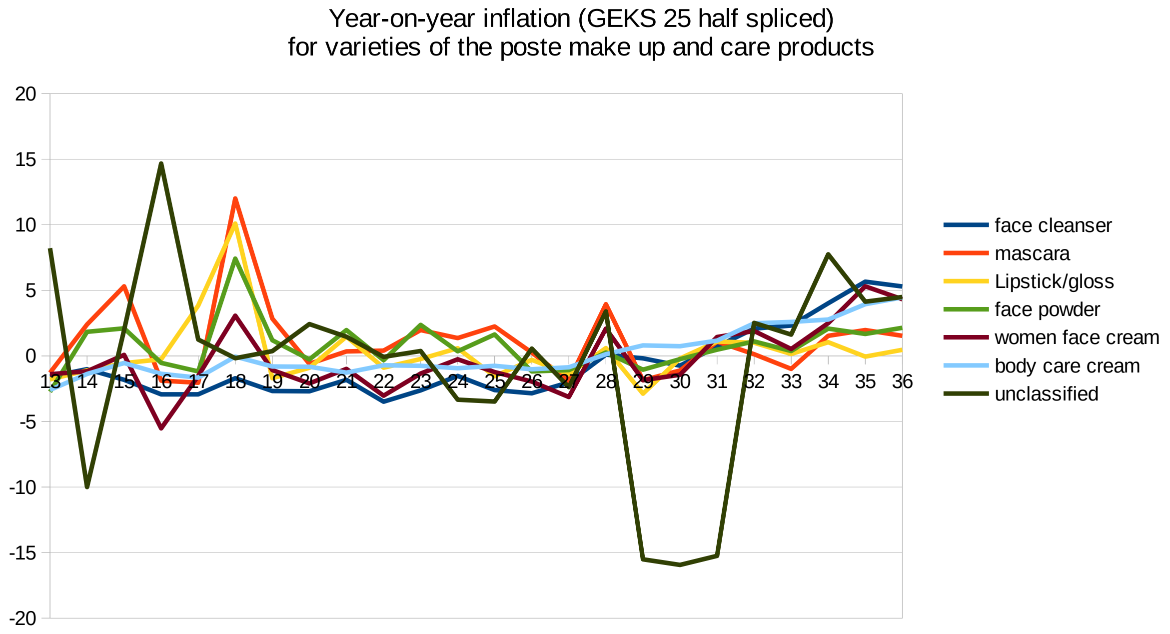

Figure 25: Figure 25: Source: scanner data. Scope: Metropolitan France. Reading note : Between February 2020 and February 2021, the price level for unclassified canned meat products has decreased of 10,0 %. Year-on-year inflation is computed as $100*(\frac{I_{m,y}}{I_{m,y-1}}-1)\%$

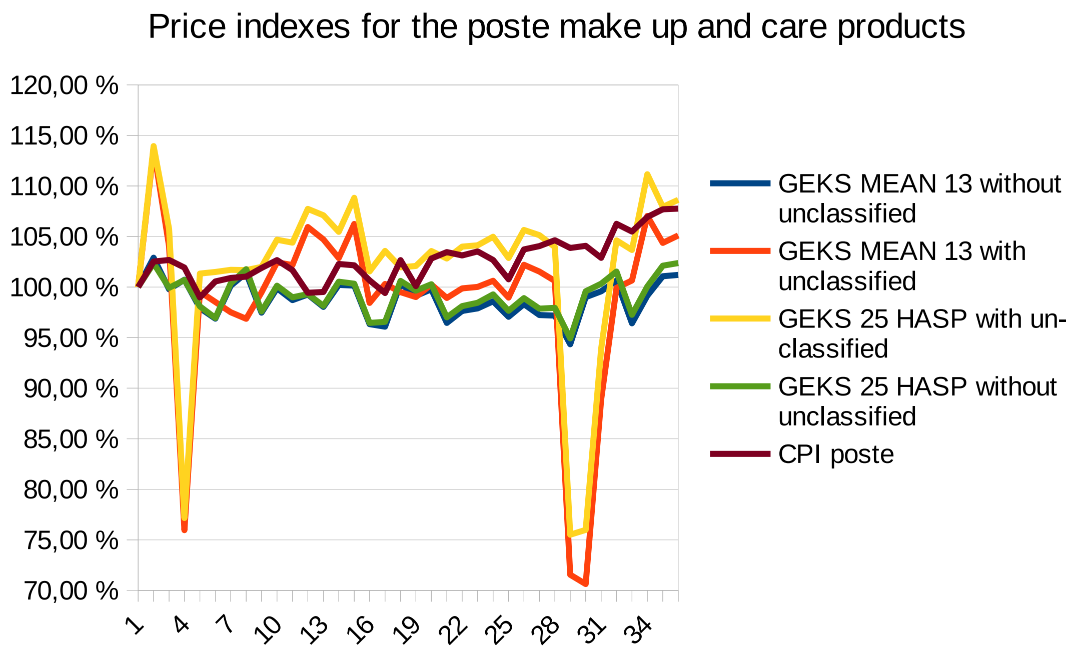

Figure 26 Source: scanner data and French CPI. Scope: Metropolitan France. Reading note: in period 12 (December 2020), GEKS index for the poste make up and care product including the unclassified data grouping by EAN x outlet with a window of 13 month and mean splicing was 105.9. The GEKS indexes are computed with a Törnqvist index formula, the splicing method is mean for the window size 13 and half for the window size 25.

As we can see just above, the unclassified data have a strong impact. Also, we can see that the window length has no real impact on the multilateral indexes if unclassified are excluded. But, regardless of this length, the multilateral indexes are quite far from the current one. This current index shows an increase of prices when the multilateral ones demonstrate more stability.

GEKS indexes by Circana family in 2020 for unclassified data

at the make up and care poste level

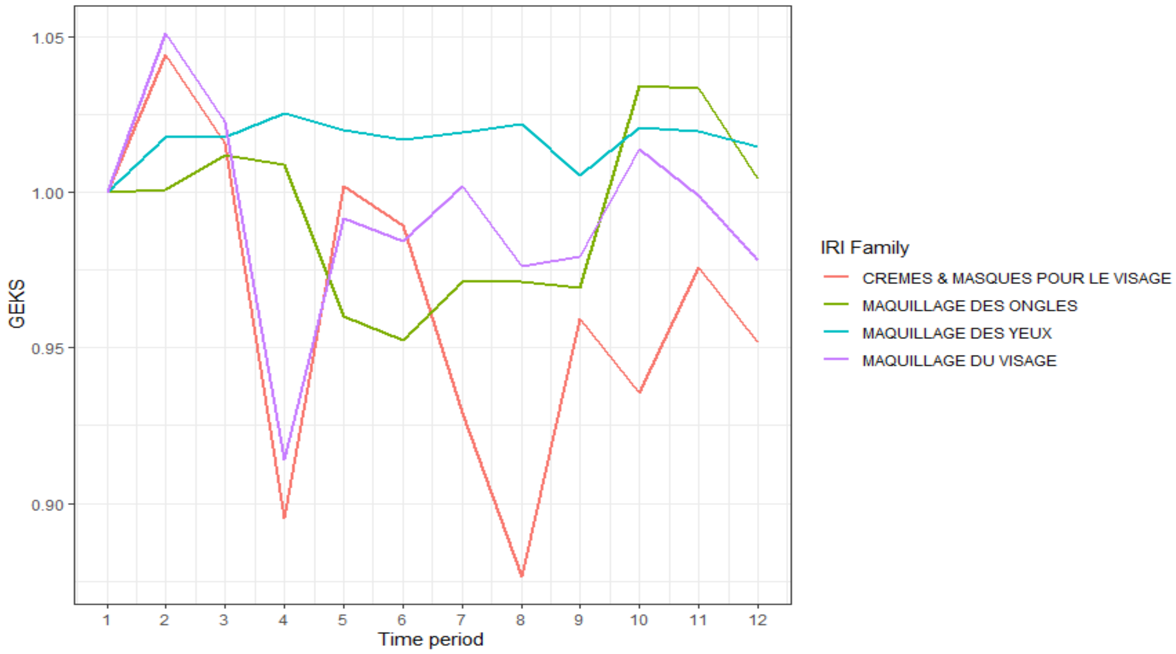

Figure 27: Source: scanner data in 2020. Scope : Metropolitan France. Reading note: in November 2020, GEKS index for the Circana (previously IRI) family "eyes make-up" grouping by EAN x outlet with a window of 25 month and half splicing was 97.5. Only the four family with the highest expenditure share are represented

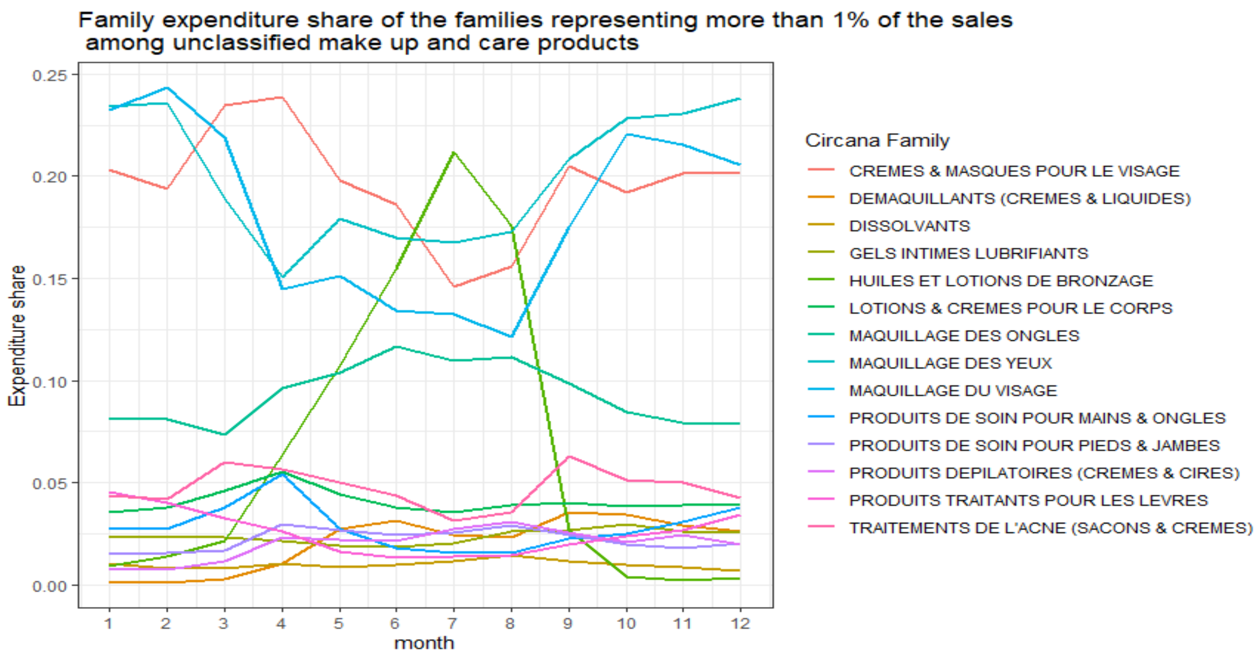

Figure 28: Source: scanner data in 2020. Scope : Metropolitan France. Reading note: in November 2020, the Circana (previously IRI) family "eyes make-up" represented 23.1% of the expenditure of unclassified make up and care products. Only families representing more than 1% of the expenditure are represented.

Among unclassified data, some are linked to Circana Families representing field collected varieties, some others don't respect the precise specifications of varieties : for instance eyes make-up isn't in our scanner data varieties currently.

4) Contributions behind GEKS-Tq variation

a) Theory

To understand better the variation of the indexes computed on the previous section of this paper, we wanted to take a look into the contributions of individual products to the index variation[6].

A contribution is defined between two periods

To do so, we have to start by looking at the contribution in the bilateral index. If the product is present at both periods, its expenditure share in the corresponding bilateral index is :

$$w_i^{t1,t2}=0,5*\left(\frac{p_i^{t1}q_i^{t1}}{\sum_{j\in N_{t1}\cap N_{t2}} p_j^{t1}q_j^{t1}}+\frac{p_i^{t2}q_i^{t2}}{\sum_{j\in N_{t1}\cap N_{t2}} p_j^{t2}q_j^{t2}}\right)$$

were p_i^t is the price of product i at period t and q_i^t is the number of product i sold at period t. It is the weight it has in the Törnqvist index.

We are then able to compute the average bilateral share of this product from a period t with all the other periods in the window in the multilateral index:

$$w_i^{*,t}=\frac{1}{card\,W}\sum_{r\in W} w_i^{r,t}$$

To look at the contribution of a product in the index variation between period t1 and t2, we have to apply these weights to the product of price variation in each period: which give the formula:

$$I_{GEKS-TQ}^{t1,t2}=\prod_{i\in N}\frac{(p_i^{t2})^{w_i^{*,t2}}}{(p_i^{t1})^{w_i^{*,t1}}}\prod_{t\in W}(p_i^t)^{\frac{w_i^{t,t1}-w_i^{t,t2}}{card\,W}}$$

and so we have a decomposition

$$I_{GEKS-TQ}^{t1,t2}=\prod_{i\in N} contribution_i^{t1,t2}$$

With this formula, the index is represented as the product of the contribution of each product. In order to facilitate the interpretation by having a summability between contributions, we looked at the log of the index and the log of the contributions.

$$\ln(I_{GEKS-TQ}^{t1,t2})=\sum_{i\in N}\ln(contribution_i^{t1,t2})$$

$$\ln(I_{GEKS-TQ}^{t1,t2})=\sum_{i\in N}\ln\left(\frac{(p_i^{t2})^{w_i^{*,t2}}}{(p_i^{t1})^{w_i^{*,t1}}}\prod_{t\in W}(p_i^t)^{\frac{w_i^{t,t1}-w_i^{t,t2}}{|W|}}\right)$$

Our product definition here, is still an EAN x Outlet.

6 Thanks to the paper, Decomposing Multilateral Price Indexes into the Contributions of Individual Commodities, and the guide on Multilateral Methods (Chapter 8) guide we looked into this direction.

Due to expensive computation costs, these contribution are computed only at a level EAN x Outlet. We had to explore ways to reduce the number of observation and time of computation.

We used this R package: https://github.com/MjStansfi/GEKSdecomp/. It seems to be compatible only with comparing the two last periods of the window. We would have to adapt this tool to have more flexible options to analyse contributions to evolution. Also, this is limited at analysis “inside” a given window. But for longer evolution, splicing has probably to be taken into account.

b) Experiment

From our previous results, several periods for each varieties/poste where interesting to look at in order understand better the indexes variation (between period 4 and 5 for lipstick and 5 and 6 28,29 et 29,30 between 1 and 2 for unclassified make up & car products, between period 1 and 12 for unclassified canned meat, between period 3 and 4 and 4 and 5 for unclassified canned meat).

For practical reasons, we chose to calculate contributions for lipstick between period 28 and 29, with a window size of 13, without separating outlets. We studied the contributions for each EAN.

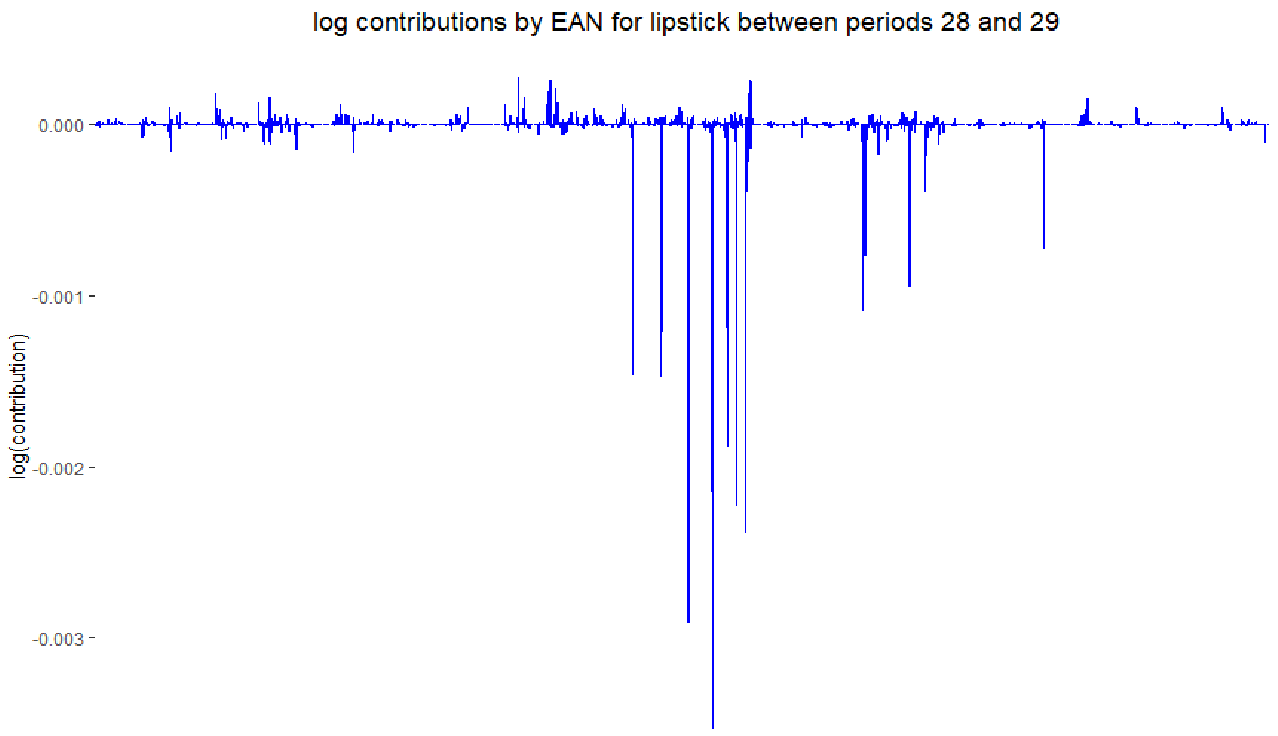

Figure 29: Source: scanner data. Reading note: Each bar represent the log contribution of an EAN in the price evolution of lipstick/gloss between period 28 (April 2022) and period 29 (May 2022), measured inside a window of 13 months. A log contribution superior than 0 means that the EAN contribute positively (price increase) and a negative one negatively.

By transitivity of the GEKS index, we can compare contributions between period 28 and 29 to the ratio of $\frac{I_{GEKS-TQ}^{17,29}}{I_{GEKS-TQ}^{17,28}}$. We cannot theoretically compare the ratio of spliced indexes.

GEKS using EAN and window size of 13 for lipstick:

Time Period	GEKS-Tq EAN 13 Mean splice	GEKS-T ean X outlet 13 Mean splice	GEKS by EAN without splicing (period 17 as reference)
17	102,70	98,90	100
18	104,01	99,05	100,88
19	100,56	97,05	96,88
20	104,60	98,74	99,94
21	104,06	97,30	98,89
22	104,25	98,00	99,55
23	103,39	98,90	100,8
24	103,31	98,33	99,66
25	94,12	96,73	95,42
26	98,87	98,22	96,77
27	105,15	96,91	97,31
28	105,90	98,90	99,38
29	104,57	95,55	96,97

$$\frac{I_{GEKS-TQ}^{17,29}}{I_{GEKS-TQ}^{17,28}} = \frac{96,97}{99,38} = 0,9756 = 97,5\%$$

With the contributions computed with the R package GEKSdecomp we have the following results:

$e^{\sum_{i \in EANs} \log(contrib_i)}$ = 0.9691 = 96,9%

We are also able to find the EAN with the contribution the furthest from 1. It has a log(contribution) of -0.00353 and a contribution of 0.9965

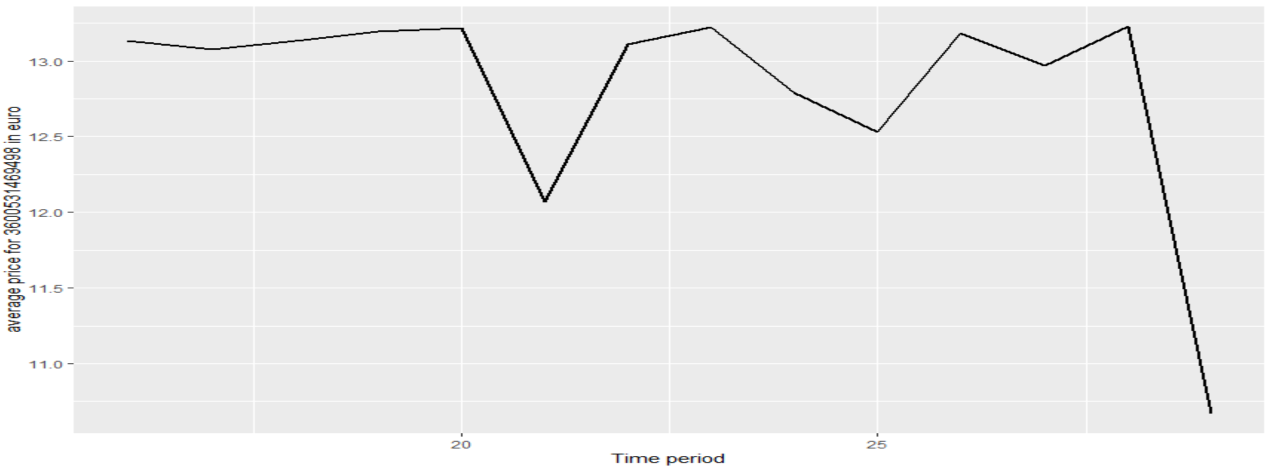

Figure 30: Source: scanner data. Scope: Metropolitan France. Reading note: in period 20 (July 2021), the average price for the EAN studied is 13.21€. The average price is computed as an average weighted by the expenditure share.

IV) Next steps ? Our « research » agenda

This work is the beginning of a longer project about multilateral methods and their interest given our context. If this work shows some interesting leads, a lot has still to be done.

1) Link between those multilateral indexes and microeconomic theory

First, while we kept some very close methodology for our scanner data, we were able to use the same explanation for our methodology. This is a fixed basket representing the mean consumption of French households optimising their utility. As some links exist between Laspeyres, Paasche and Fisher indexes on one hand and micro-economic theory one the other hand: there is some theoretical grounds to our current method.

At this stage, we need to better understand the economical approach on which are based multilateral methods and how to communicate and interpret results with these methods. This is of interest to make this index understandable by anyone in society.

2) Explore the outlet dimension

In our current CPI methodology for field collected varieties, we use sampling and define targets among the outlet according to their classification (supermarket, hypermarket, specialized shop). For scanner data varieties, they represent only two kind of outlet: supermarket and hypermarket. In this experiment, we are producing micro indexes at the outlet index, which means that we consider for the customer there is no substitution between buying in a shop or another. The latter point can be discussed, because for instance we could consider that outlets of the same size, from the same retailer and in a close geographic area could be considered equivalent. Following ean into a group of shop could improve the quality of the index because it could improve the match rate between periods.

3) Going further with classification methods

As explained above, this work requires a classifying tool. Without this, we cannot classify data into the COICOP and consequently, we cannot compute relevant indexes. This task will be tackle in the following months by making progress with the existing tools we have.

We use a fasttext algorithm which is a neural network tool specialized in dealing with characters strings. By extracting labels of products and their corresponding expenditures we will optimize the classifying function. Our goal is to have a good performance at the “poste” level – going further seems to be unreasonable given the information we have.

4) Strategy to include those indexes inside our current methodology

Before hoping to use these indexes in production, we have to deepen our look into the contributions, the interpretability/decomposition of an index evolution. We presented some first contribution computation but we will need to conceive more practical routines for understanding such index evolutions.

And when we will be able to classify product, to compute multilateral index with enough understanding of it (from both statistical and theoretical approaches), we will have to have a reflection on how it will be possible to use this kind of methods with the rest of the basket we follow and to see how to adapt this with our current methodology (whether to change everything or to have some cohabitation).

V) Conclusion

This first real experimentation of multilateral index with our scanner data gives us some first learnings :

- at a really fine scale, this index behaves quite closely to our current methodology and consequently seems possible to work at the EAN scale – it sill remain to be confirmed at a larger scale
- working at the EAN scale seems to be acceptable but it has to be confirmed with a larger scale experiment
- at a more aggregate scale (our "poste level"), there is more volatility and we have to progress in our understanding and tools for this

This first work emphasises two major elements that we need to work on :

- Classification tool to classify the products in the COICOP nomenclature
- Theoretical understanding of the links with micro-economic theory

References

- Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Eurostat, 2022.
- MARS: A method for defining products and linking barcodes of item relaunches, Antonio G. Chessa, Statistics Netherlands.
- “Chain drift” in the Chained Consumer Price Index: 1999–2017, Monthly Labor Review, BLS December 2021.
- Évaluation des méthodes multilatérales de calcul de l'indice, STATBEL, Ken Van Loon et Dorien Roels, 07/2019
- Eliminating Chain Drift in Price Indexes Based on Scanner Data, Jan de Haana and Heymerik van der Grient, Statistics Netherlands,2 April 2009
- FMI, CPI Manual, 2020.
- From GEKS to cycle method , 11/2017, Leon Willenborg
- A Closer Look at the Rolling Window GEKS Index with a Movement Splice, Jan De Haan, 16 October 2017
- Extension of multilateral index series over time: Analysis and comparison of methods, Antonio G. Chessa, 7 May 2021
- Transitivity of price indexes, Leon Willenborg , May 2018
- Comparing Price indexes of Clothing and Footwear for Scanner Data and Web Scraped Data
- Antonio G. Chessa* and Robert Griffioen**, *Statistics Netherlands, Team CPI* ,1st April 2019
- Leclair (2019), « Utiliser les données de caisses pour le calcul de l’indice des prix à la consommation », Le Courrier des statistiques, n°3
- Decomposing Multilateral Price Indexes into the Contributions of Individual Commodities, Michaël Webster and Rory C. Tarnow-Mordy , 2019
- Introducing multilateral index methods into consumer price statistics, Liam Greenhough , ONS, 28 November 2022
- The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity, Caves, Christensen and Diewert, 1982
- The use of weighted GEKS for the calculation of consumer price indexes: an experimental application to Italian scanner data Alessandro Brunetti (Istat), Stefania Fatello (Istat), Tiziana Laureti (Università della Tuscia), Federico Polidoro (Istat) *17th Ottawa Group Meeting, Rome, 7 – 10 June 2022*

Appendix

Average price evolution of Circana families among unclassified make up and care products between January 2020 and December 2022

Figure 31: Source: Scanner data. Scope: Metropolitan France. Reading note : The evolution is computed as the ratio of average price at period m/ average price at period 0

Expenditure share of Circana families among unclassified make up and care products between January 2020 and December 2022

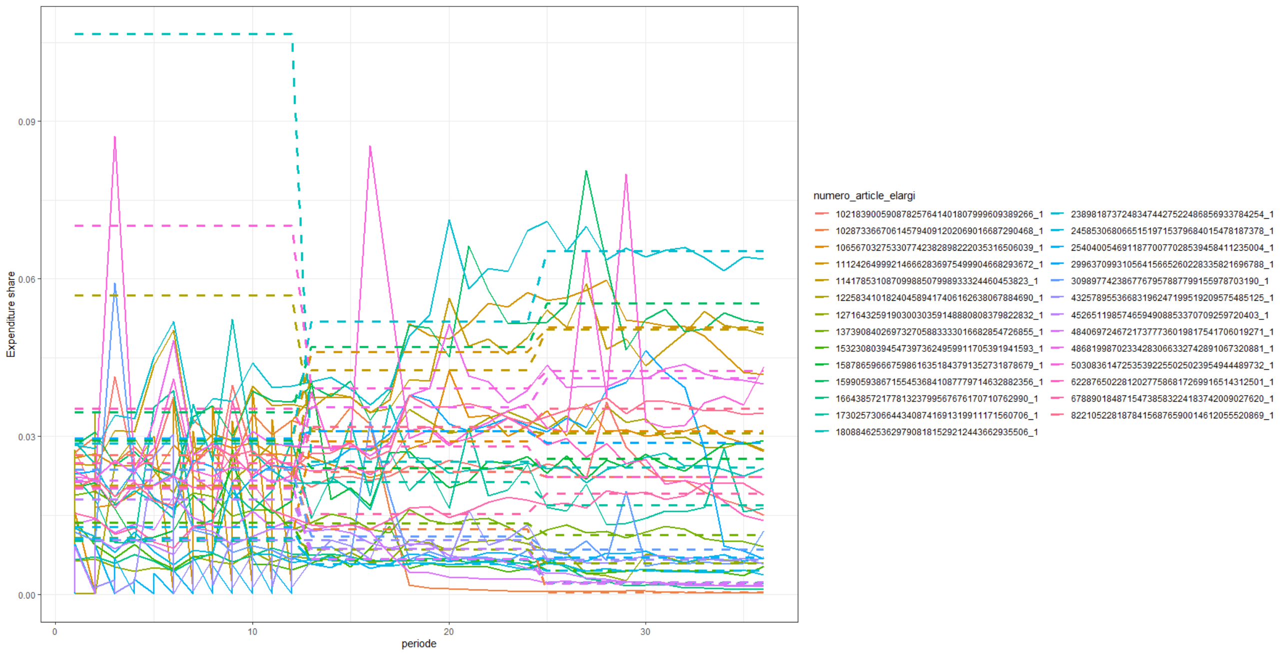

Figure 31: Source: Scanner data. Scope: Metropolitan France. The dotted lines represent annual expenditure shares and the continuous one monthly shares. Results are presented for extended group representing more than 1 % of the expenditure in 2020