Exploring methodologies to integrate new scanner data in the French CPI:

Making use of multilateral methods

Insee

Measuring, understanding

1 CONTEXT AND GOALS

2 THEORY

3 RESULTS : BY VARIETY (COICOP 7 DIGITS)

4 RESULTS : BY COICOP 6 DIGITS(MAKE UP)

5 RESULTS : CONTRIBUTIONS

01 INTRODUCTION

− **We are starting to receive data from 2 hard discounters.**
− **We already have and use in production (since Jan 2020) scanner data from other retailers**
− **Our current methodology with scanner data requires an external referential allowing us from GTIN/EAN to have**

- Additional characteristics (volume, unit, label, color ...)
- Nomenclature
  − With classification rules and using the characteristics we classify at the variety level (level 7 of COICOP, French specificity).
- We are able to group EAN into equivalence classes to follow products better, avoid basket churn and catch the relaunches.
- We compute a Geometric Laspeyres, the methodology is similar than with the field collected data and the quality adjustment is slightly different since we can use the price history for the replacement product.

− **Hard discount data has for now a low match rate with the referential (17 % of expenditure share according to 1 test file for one retailer and 39% for the other)**

− **We will experiment multilateral methods mainly to check what we could do without the referential and with the constraints of avoiding chain drift and basket churn.**

- **We will use our already possess scanner data (not enough history with hard discounters)**

- **Our product definition will vary between using GTIN/EAN or a article grouping methods (extended article number)**

- **We will compute micro indexes at the outlet level.**

- **We follow the average price of each product per month.**

- **Scanner data from January 2020 to December 2022, from 6 retailers (without hard discount because we don't have background data).**

- **3 varieties & their corresponding 6 digits COICOP level**
  - **Whole milk & whole milk=> few replacements**
  - **Foie gras & canned meat=> a high seasonality and 85% of replacement during the year**
  - **Lipstick & make up and care products => a lot of distinct GTIN/EAN.**

02 MULTILATERAL METHODS

– **We focus on GEKS-Törnqvist**

$$I_{GEKS}^{0,t} = \prod_{l=0}^{T} \left(\frac{I^{0,l}}{I^{t,l}} \right)^{\frac{1}{T+1}} = \prod_{l=0}^{T} \left(I^{0,l} * I^{l,t} \right)^{\frac{1}{T+1}}$$

where

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

and

$$s_i^t = \frac{p_i^t q_i^t}{\sum_{j \in S} p_j^t q_j^t}$$

– **The sample S can be a COICOP 6 digit level or a variety**

– **The product i can be the GTIN/EAN or an Extended article number**

- **Choice of the window size and splicing:**
  - **Rolling window of size 13 and mean splice**
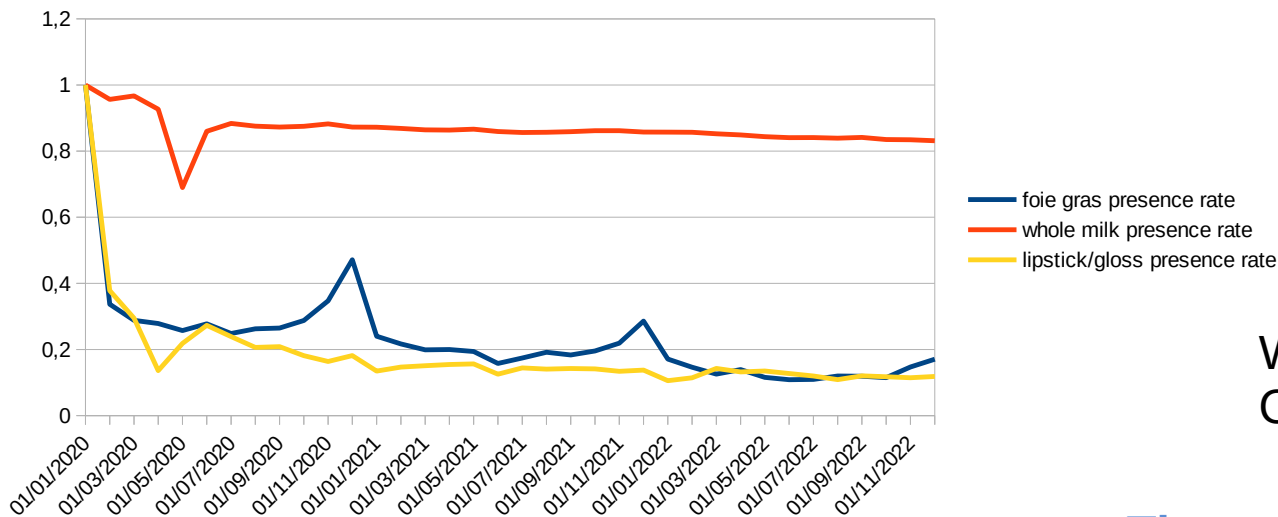 - **Rolling window of size 25 and half splice**

– **Using R and IndexNumR package**

03 RESULTS : VARIETY

Proportion of EAN x Outlet present in January 2020 and at the month m

- foie gras presence rate
- whole milk presence rate
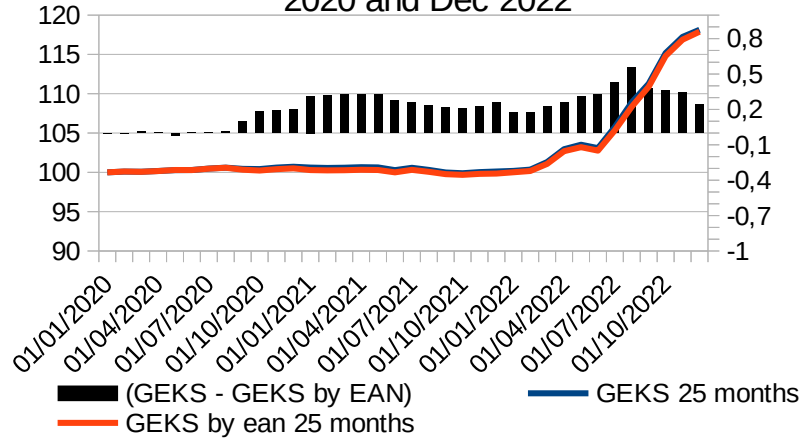- lipstick/gloss presence rate

The presence rate is computed as
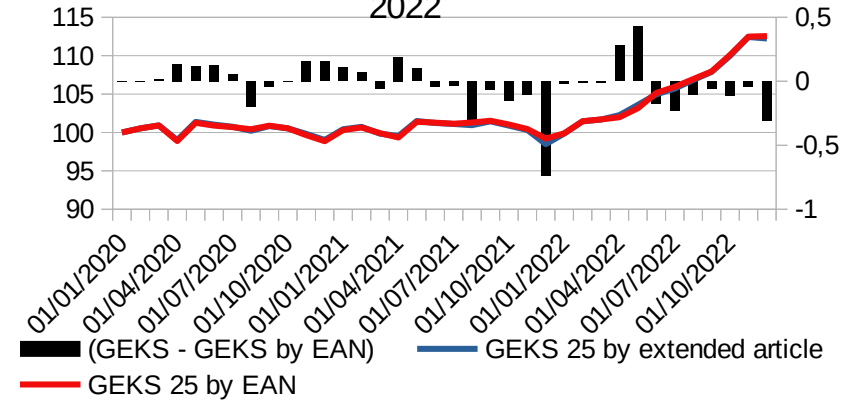
$$\frac{|N_i \cap N_1|}{|N_1|}$$

Where N_i are the products (EAN X Outlet in our case) sold in period i.

– **The presence rate is low for foie gras and lipstick**
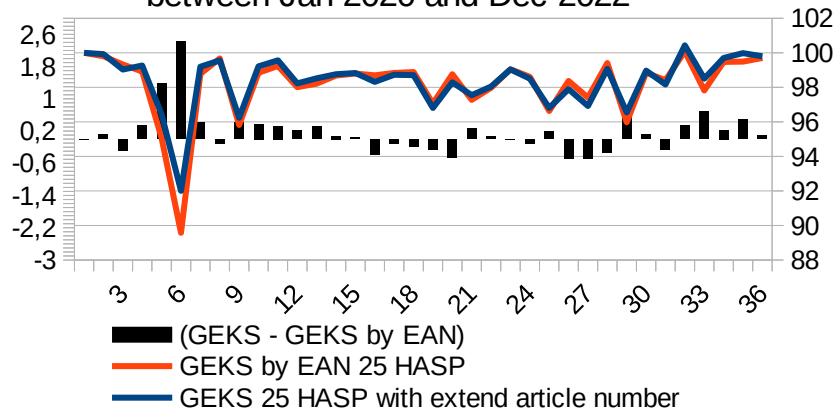
– **There is a seasonality for foie gras**

Price indices for the variety whole milk between Jan 2020 and Dec 2022

- (GEKS - GEKS by EAN)
- GEKS 25 months
- GEKS by ean 25 months

Price indices for the variety foie gras between January 2020 and December 2022

- (GEKS - GEKS by EAN)
- GEKS 25 by extended article
- GEKS 25 by EAN

Price indices for the variety lipstick/gloss between Jan 2020 and Dec 2022

- (GEKS - GEKS by EAN)
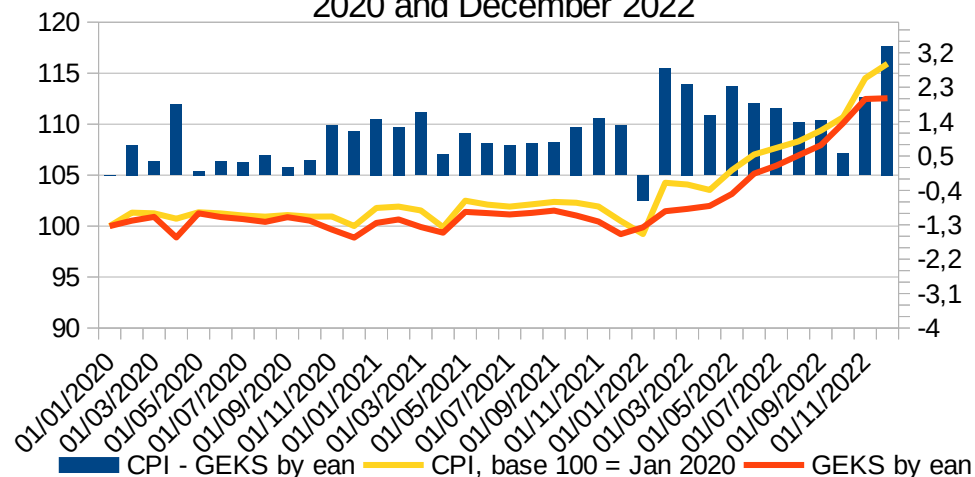- GEKS by EAN 25 HASP
- GEKS 25 HASP with extend article number

Indexes using EAN or extended article number are very close for milk and foie gras.
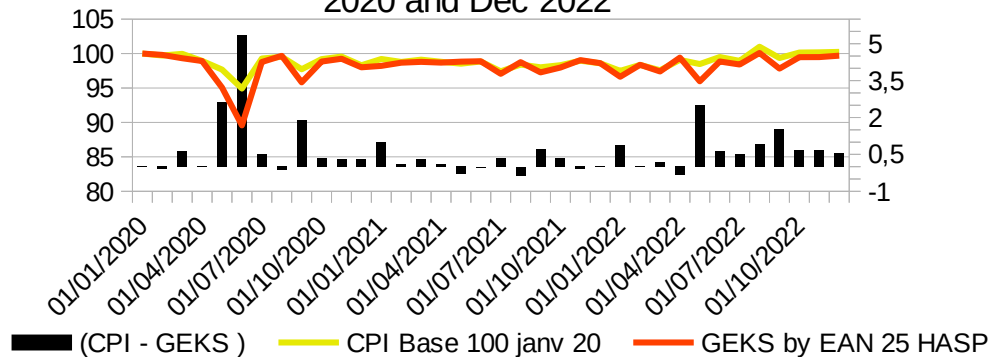
There is more volatility for lipstick.
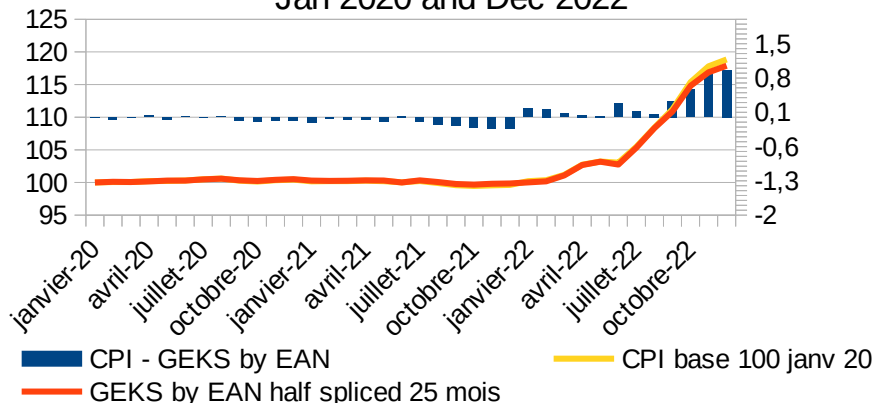
Insee — Measuring, understanding

Price indices for the variety whole milk between Jan 2020 and Dec 2022

Legend:
- CPI - GEKS by EAN
- CPI base 100 janv 20
- GEKS by EAN half spliced 25 mois

Price indices for the variety foie gras between January 2020 and December 2022

Legend:
- CPI - GEKS by ean
- CPI, base 100 = Jan 2020
- GEKS by ean

Price indices for the variety lipstick/gloss between Jan 2020 and Dec 2022

Legend:
- (CPI - GEKS)
- CPI Base 100 janv 20
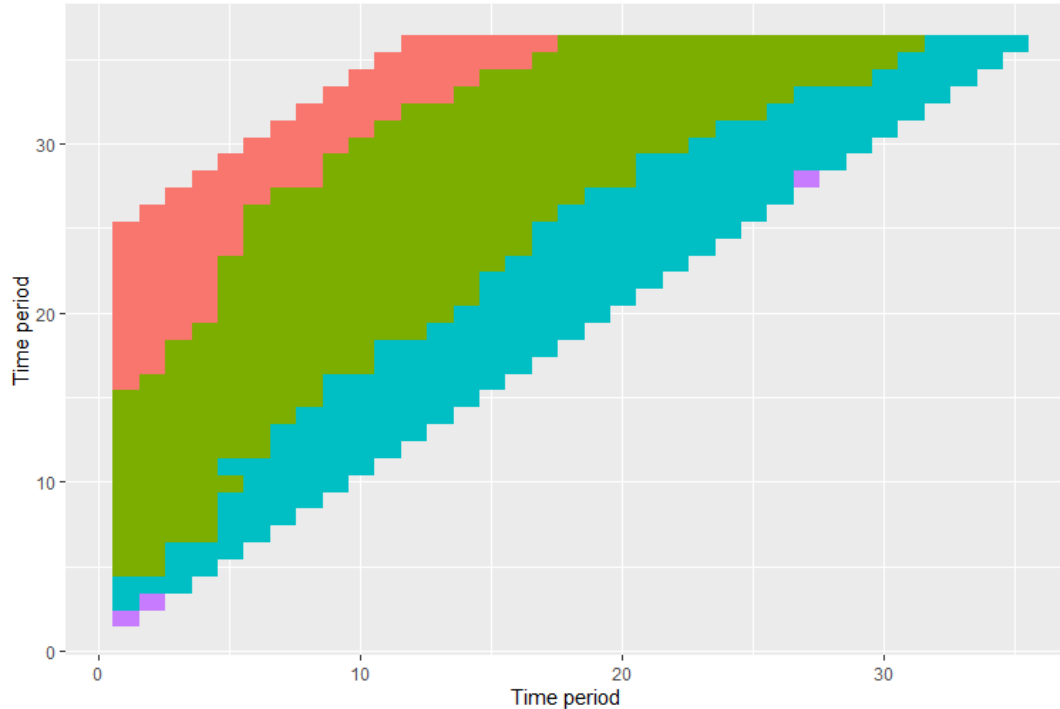- GEKS by EAN 25 HASP

There is more difference between GEKS and CPI than between two GEKS.

A highest volatility in some period (June 2020 for lipstick for instance)

04 RESULTS: COICOP 6 DIGITS (MAKE UP)

Ean X outlet match rate between two time periods within a window of 25 months for make up and care products (without unclassified).

Match rate
- (0.2,0.3]
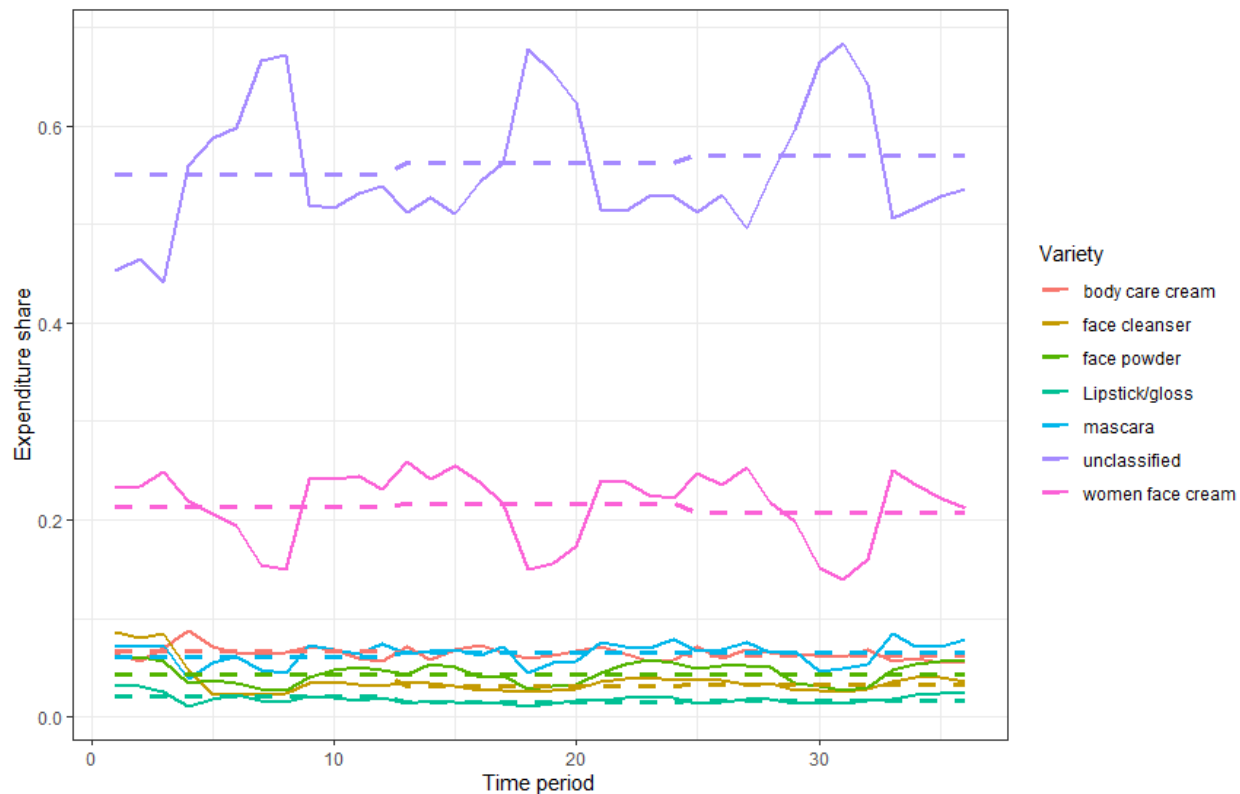- (0.3,0.4]
- (0.4,0.5]
- (0.5,0.6]

The match rate is computed as

$$\frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

Where N_i **are the products (EAN X Outlet in our case) sold in period I.**

There might be lockdowns effects in some periods.

Even for two consecutive periods, the match rate is quite low.

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


- **In some COICOP 6 digits level we have a high proportion of unclassified data they can be**
  - Linked to field varieties (we do not have yet a corresponding scanner data variety) : nail make up for instance
  - Do not correspond to the classification rules (a canned meat with honey flavour for instance)

They aren't followed taken into account our current CPI.

- **What would be the impact of keeping them ?**

Year-on-year inflation (GEKS 25 half spliced) for varieties of the poste make up and care products between Jan 2021 and Dec 2022
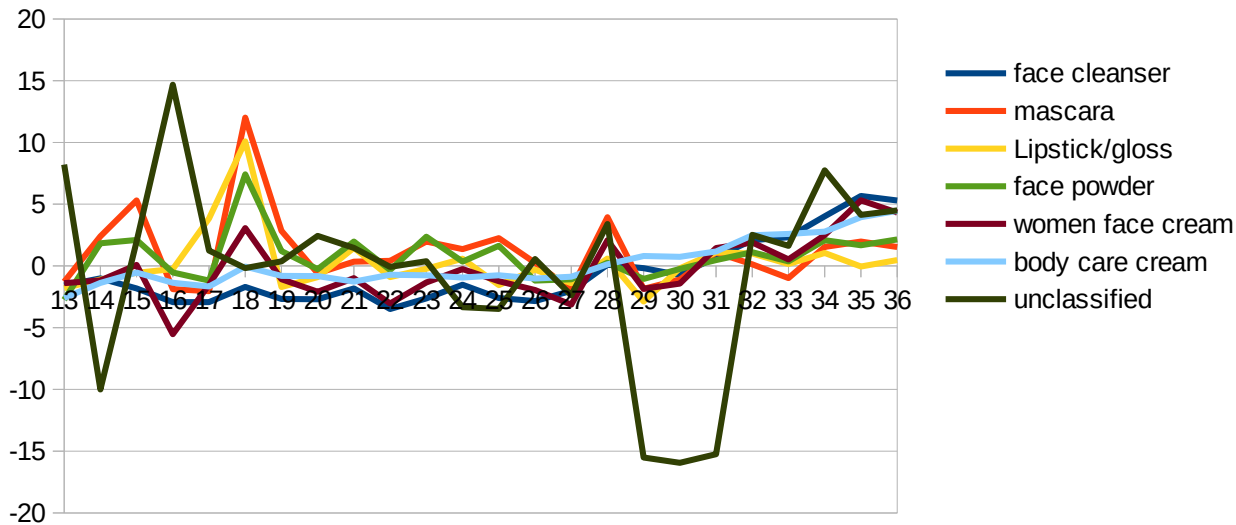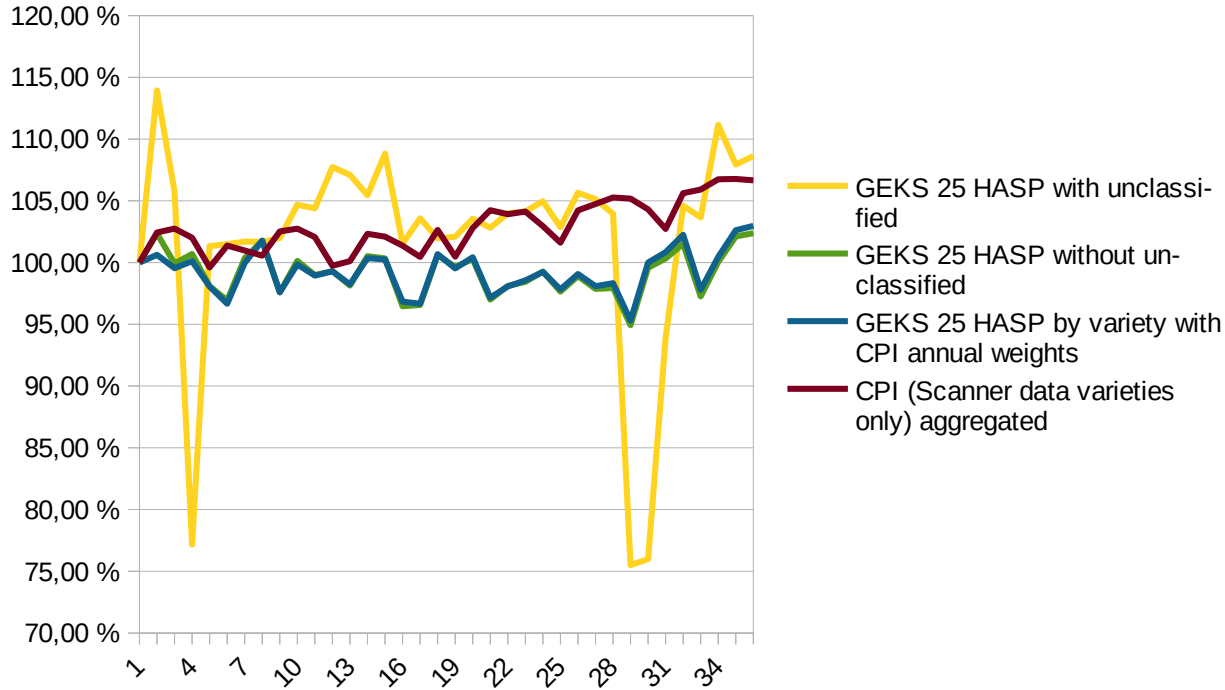
- **The index for unclassified data is more volatile**
  - **Beginning of 2021**
  - **Summer 2022**

Price indexes for the poste make up and care products

The trend is kept with unclassified data but there is a high volatility

05 RESULTS : CONTRIBUTIONS

- **Goal : understand and explain the index variation from the observations**

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

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

with unspliced indexes we can use the transitivity : $\quad I_{GEKS-TQ}^{12,13} = \dfrac{I_{GEKS-TQ}^{1,13}}{I_{GEKS-TQ}^{1,12}}$

- **We could also make sub groups (retailer, geo) to sum log(contrib)**

Using R and GEKSDecomp package

log contributions by EAN for lipstick between periods 28 and 29

GEKS-Tq ean X out-let 13 Mean splice

GEKS-Tq by EAN wi-thout splicing (period 17 as reference)
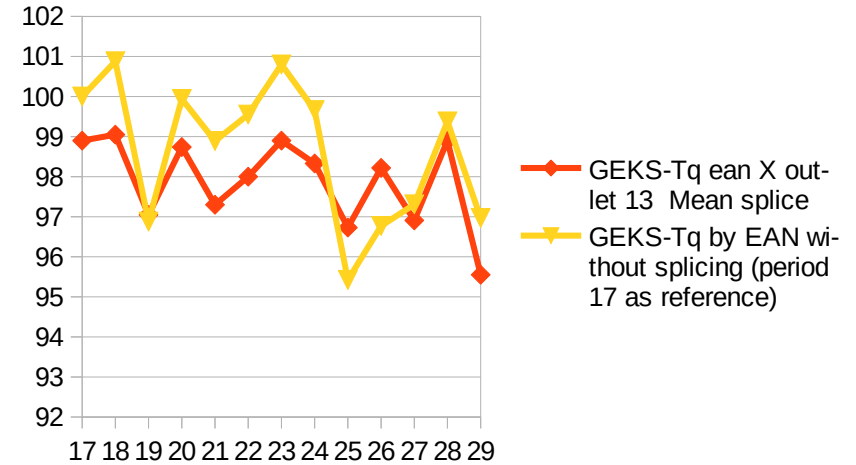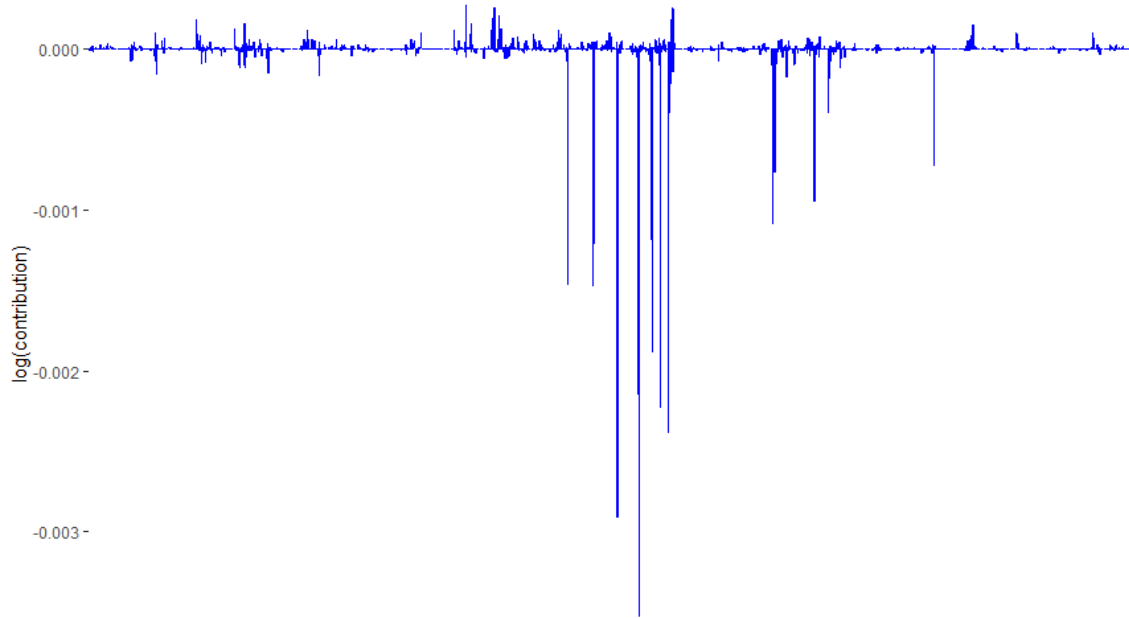
The decrease of the index between period 28 (June 2022) and 29 is carried by few products

- **Learnings**
  - At a really fine scale (GTIN/EAN), the GEKS indexes behave quite closely to our current methodology
  - At a more aggregate scale, there is more volatility and we have to progress in our understanding and tools including classification issues.

- **Future works**
  - Classification tools
  - Theoretical understanding of the link with micro-economic theory

Join us on

insee.fr

Adrien Montbroussous & Martin Monziols
Methodologist & Head of the methodology unit
Consumer Prices Division
adrien.montbroussous@insee.fr
martin.monziols@insee.fr

Measuring, understanding

Insee