New data sources : scanner data

The case of scanner data for the French consumer price index

Insee

Measuring, understanding

Measuring, understanding

01 SCANNER DATA: DEFINITION

SCANNER DATA ARE DATA COLLECTED BY RETAILERS WHEN CONSUMERS PAY FOR GOODS IN STORE. THE BARCODES (EAN/GTIN) OF PURCHASED GOODS ARE RECORDED, TOGETHER WITH, FOR EACH BARCODE, THE PRICE AND QUANTITY OF GOODS PURCHASED.

EXTRACT FROM SCANNER DATA USED BY FRENCH CPI

4 011200 296908 >
country
company
article
Control key

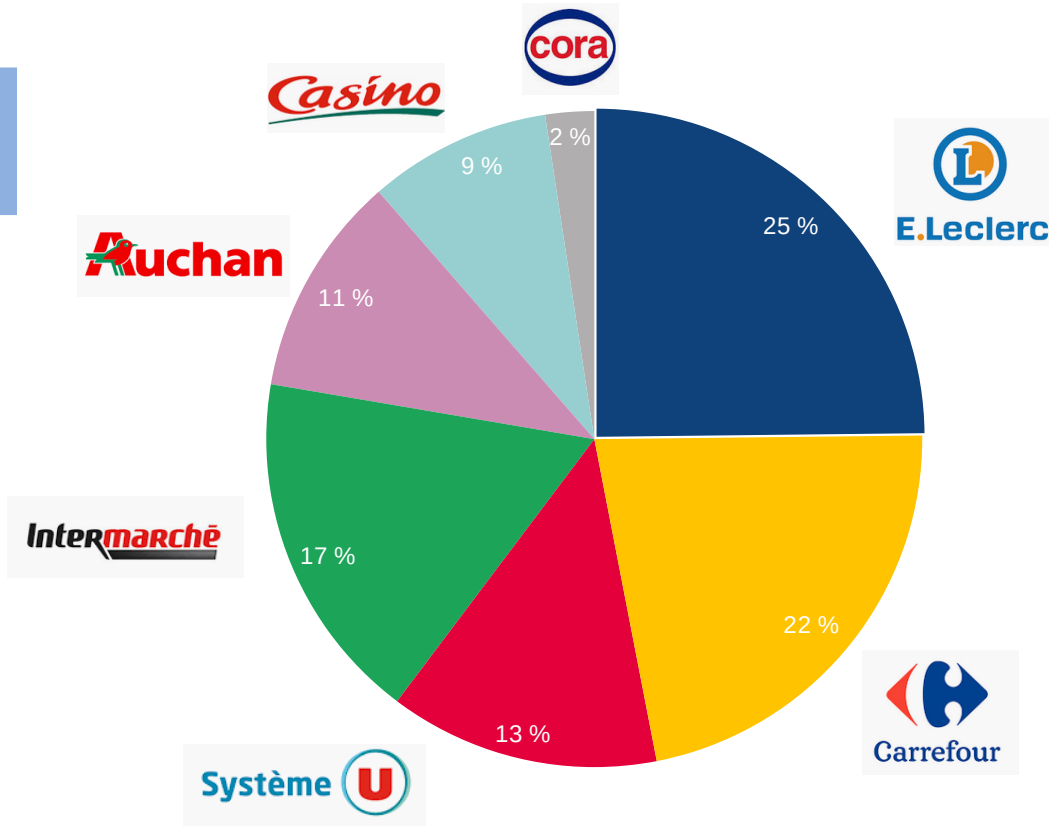| ID OUTLET	EAN/GTIN	DESCRIPTION	DATE OF SALE	QUANTITY	PRICE (€)	TURNOVER (€)
933	3272770004817	ST MORET PLAIN 150G	20200108	10	1.89	18.90
933	3154230040286	HERTA BACON 150G	20200108	7	2.38	16,66
825	3184670001080	RIANS STRAINED SOFT 6%MG 1KG	20200128	25	2.59	64,75

Insee Measuring, understanding

Distribution of retailers' turnover in the « scanner data basket »

How to obtain data ?
Law versus (or in addition to) agreements: the digital law.

Received data :
Daily transmission of scanner data (for each item and each outlet), two days after the sales.
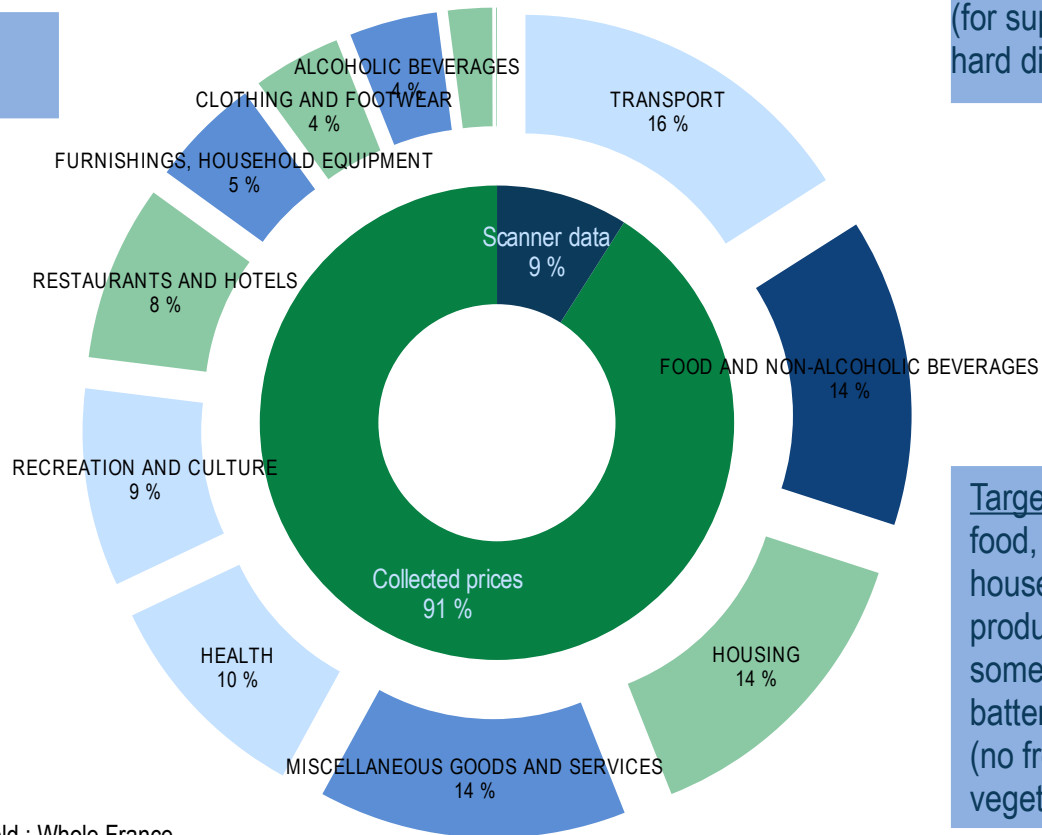

cora — 2 %
Casino — 9 %
E.Leclerc — 25 %
Auchan — 11 %
Carrefour — 22 %
Intermarché — 17 %
Système U — 13 %

Consistent with external statistics (Kantar)

Field : Mainland France
Source : Scanner Data, Insee treatments

02 USE FOR FRENCH CPI

Distribution of french household consumption (in %) and scanner data field

9% of the CPI basket comes from the scanner data

Scanner data from all of the big retailers (for super and hypermarkets, excluding hard discount)

Chart labels:
- Scanner data 9 %
- Collected prices 91 %
- ALCOHOLIC BEVERAGES 4 %
- CLOTHING AND FOOTWEAR 4 %
- FURNISHINGS, HOUSEHOLD EQUIPMENT 5 %
- RESTAURANTS AND HOTELS 8 %
- RECREATION AND CULTURE 9 %
- HEALTH 10 %
- MISCELLANEOUS GOODS AND SERVICES 14 %
- HOUSING 14 %
- FOOD AND NON-ALCOHOLIC BEVERAGES 14 %
- TRANSPORT 16 %

Legend:
- High impact of scanner data
- Moderate impact of scanner data
- Little impact of scanner data
- No impact of scanner data

Target products: manufactured food, beverages, household goods, pets products, products for personal care, some other products (eg. car batteries, bandages) (no fresh products : meat, fish, vegetable, fruits)

Field : Whole France
Source: CPI, weights 2020 , Insee

Scanner data : an additional source of information

➡ **No change to the core concepts of the French CPI**

- Fixed basket (excluding seasonal items) of products sold the previous year (with knowledge of weights at a fine level)

- GTIN : not a good product identifier ==> notion of "expanded article" (deal whith relaunches)

- Replacement by randomized choice in the same consumption segment and same outlet

- Quality adjustment with overlap method ; no imputation (we know the price of replacement production in m-2)

- Micro-indices at consumption segment x outlet level are computed using geometric Laspeyres

- At higher levels of aggregation, Laspeyres arithmetic

HOW TO FOLLOW PRODUCT ? HOW TO CLASSIFY ?

- **The GTIN is not a good product identifier**
  - A GTIN is not used for two different products, but two identical products for the consumer may have different GTINS
  - The GTIN does not allow to process the case of relaunches or the promotion
    - Notion of « equivalence classes » (OR expanded product) with variation around a reference volume (defined by the best-selling article)

- **The necessity of classification of the product in the COICOP nomenclature**
  - Study of previous turnover of the articles (EAN/GTIN) by experts in the relevant consumption sector.
    - 2020 : almost 600 consumption segment / variety

- **Use of GTIN repository purchased from a panelist to identify products based on their characteristics**
  - Classification of scanner data and replacements are done automatically via an application using Big Data technologies

Example with family articles 0315 - HIGHLIGHTERS

Interest variables : marque|type_de_produit|emballage|info_label|rechargeabilite|couleur| format|type_d_encre|type_de_surligneur

FAMILLE IRI	DESCRIPTION FAMILLE	EAN	MARQUE	TYPE DE PRODUIT	EMBALLAGE	RECHARGEABILITE	COULEUR	FORMAT	TYPE D ENCRE	TYPE DE SURLIGNEUR	VOLUME TOTAL	NOMBRE UNITES DU PACK	VOLUME PAR UNITE
0315	SURLIGNEURS	3086123255586	BIC HIGHLIGHTER	SURLIGNEUR	BLISTER	JETABLE	MULTICOULEURS	STANDARD	INDELEBILE	SURLIGNEUR	5CT	1CT	5CT
0315	SURLIGNEURS	3086123331655	BIC HIGHLIGHTER	SURLIGNEUR	BLISTER	JETABLE	MULTICOULEURS	STANDARD	INDELEBILE	SURLIGNEUR	5CT	1CT	5CT
0315	SURLIGNEURS	3086123546165	BIC HIGHLIGHTER	SURLIGNEUR	BLISTER	JETABLE	MULTICOULEURS	STANDARD	INDELEBILE	SURLIGNEUR	8CT	1CT	8CT

For these 3 GTINS, a single identification number in scanner data (expanded product)

LEVEL 0 : CODE 01 – MISCELLANEOUS GOOD AND SERVICE

⬇

LEVEL 1 : CODE 12.1 – Body care

⬇

LEVEL 2 : CODE : 12.1.3 - Other devices, articles and products for personal care

⬇

LEVEL 3 : CODE 12.1.3.2 - Personal hygiene, well-being and beauty products

⬇

LEVEL 6 (french specificity) « Sub-sub class »: 12.1.3.2.3 - Soaps, toothpastes and toiletries
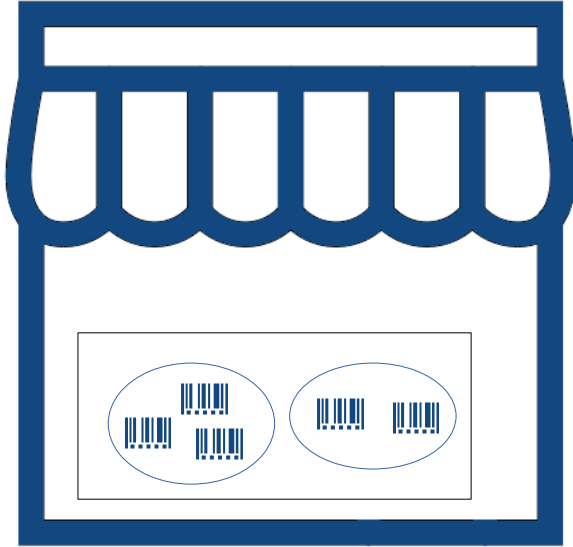
⬇

LEVEL 7 (french specificity) « Consumption segment »: 12.1.3.2.3.0003 - Liquid soap in pump bottle
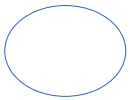
⬆

GTINs = packaging (pump bottle);
state (cream,mousse,gel,liquid);type_of_product(soap)

Outlet

GTIN

Expanded product/equivalence class

Variety

IF NEW GTIN ON THE MARKET:

- Integration in an existing equivalence class

- OR creation of a new equivalence class

  Tracking of the average price of the equivalence class at the elementary index level (id. Replacement in equivalent)

IF A PRODUCT DISAPPEARS (EQUIVALENCE CLASS X POINT OF SALE), REPLACEMENT PROCESS: :

1) Search for another expanded item in the same variety at the same outlet

2) If fails (variety that has disappeared from the outlet), search for an expanded item in the same variety at outlet in the same region.

   Quality adjustment calculated by overlap and not by imputation because we know the price in m-2 of the replacement

CPI

Laspeyres-type index

Aggregation at sub-sub class (6-digit ECOICOP index)

Consumption segment

higher level indices = arithmetic Laspeyres of elementary aggregates

$$I_{sc,m} = \sum_{outlet} w_{sc,outlet}\, I_{sc,outlet,m}$$

Consuption segment level (level 7 of the ecoicop)

Consumption segment x urban area size

Elementary aggregate : consumption segment (cs) x outlet

$$I_{cs,outlet}^{m} = \prod_{product \in cs,outlet} (IE)_{p}^{\widetilde{T/O}/\sum_{p}\widetilde{T/O}}$$

Substitution of consumer in the same outlet

Consumption segment x urban area

Unit price (by month) : EQ x Outlet = product

$$PU_{product} = \frac{\displaystyle\sum_{d \in m,\, gtin\, x\, outlet \in product} T/O_{gtin,outlet}^{d}}{\displaystyle\sum_{d \in m,\, gtin\, x\, outlet \in product} QTY_{gtin,outlet}^{d} \times VOL_{gtin}}$$

$$IE_{product}^{m} = \frac{PU_{product}^{m}}{PU_{product}^{b}} * 100$$
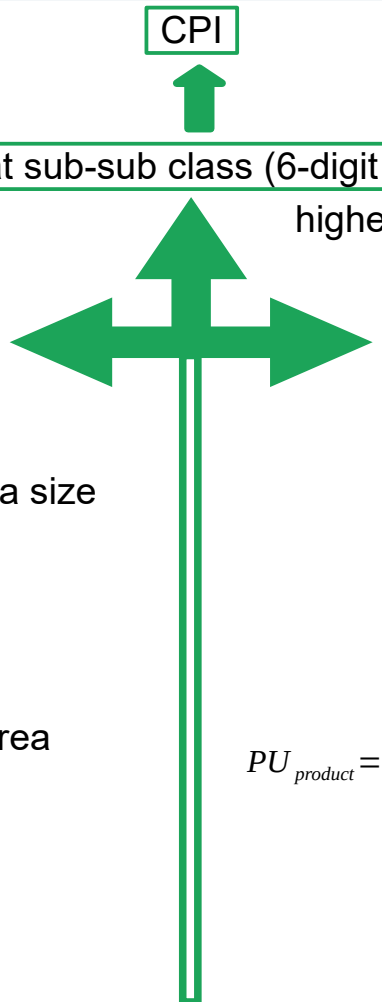
Collected price
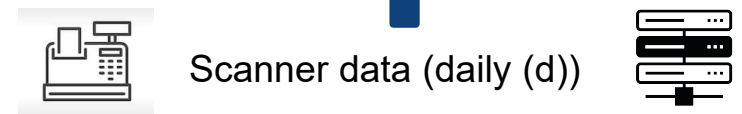
Scanner data (daily (d))

03 OTHER USES

TO PRODUCE NEW STATISTICS WITH THE SCANNER DATA :

- **Average prices**

- **Spatial comparison**

- **CPI by region on scanner data scope**

- **… and also on other area: turnover statistics, PPP**

TO OBTAIN AND EXPLOIT NEW SCANNER DATA (LIKE HARD DISCOUNTERS)

TO EXPLOIT SCANNER DATA ON NEW SCOPES OF PRODUCTS: SUSTAINABLE GOODS, CLOTHES

Thanks for your attention, any comments ?

insee.fr

Gwendoline VOLAT

Administrator of scanner data application

DG/DPCEM/IPC

+33 (0)1 87 69 63 17

gwendoline.volat@insee.fr

Insee

Measuring, understanding