TALL SOM FORTELLER

Methods for incorporating transaction data into price statistics

STATISTICS SWEDEN AND ÖREBRO UNIVERSITY SUMMER SCHOOL 2026

KJERSTI NYBORG HOV

HKJ@SSB.NO

Statistisk sentralbyrå
Statistics Norway

Alternative data sources..

..are secondary data sources that needs to be transformed to fit with the needs of price statistics.

Photo by Ivan Yerokhin on Unsplash

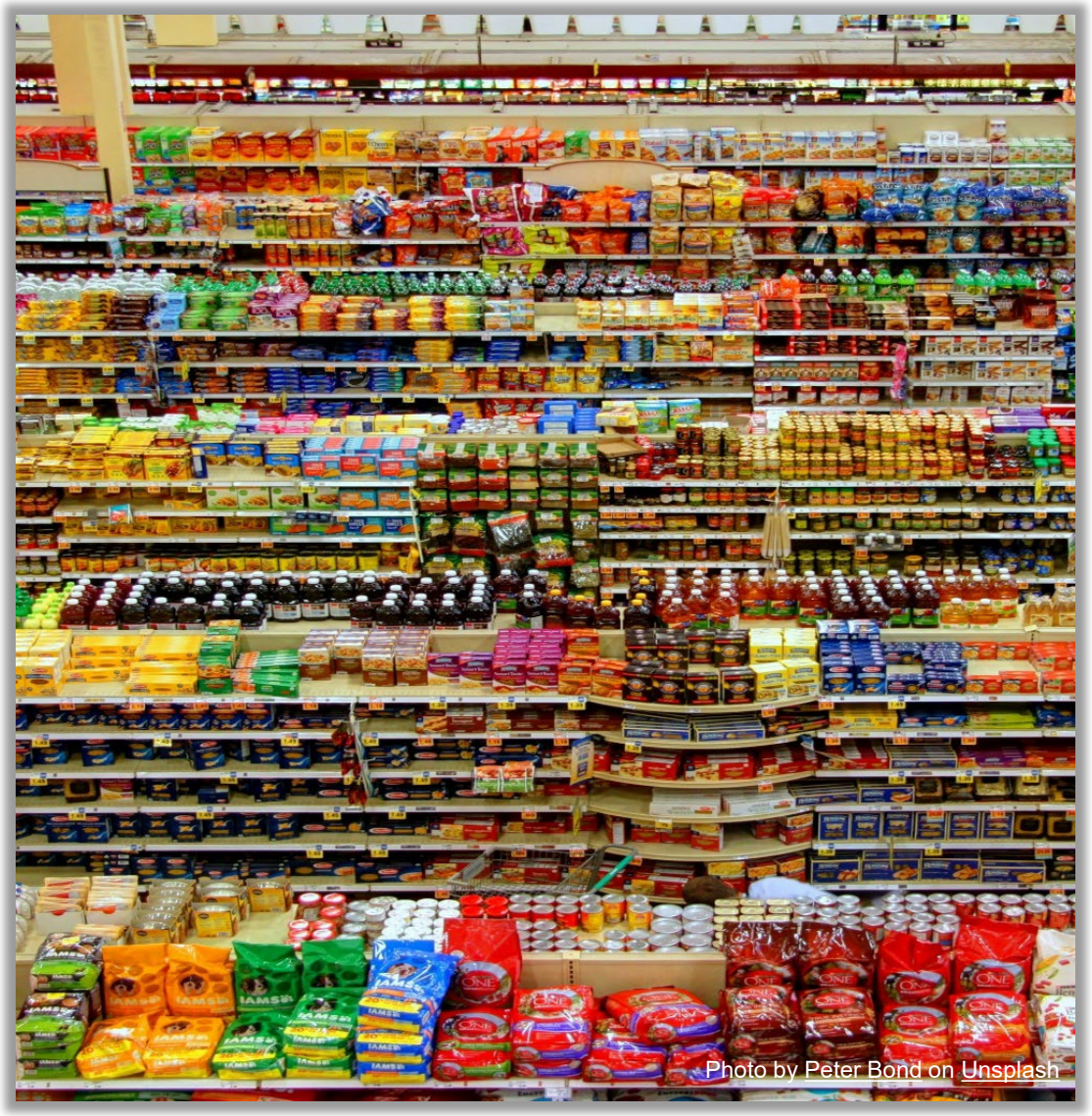


Photo by Peter Bond on Unsplash

Statistisk sentralbyrå
Statistics Norway

Data sources

- Traditional data collection
  - Representative item – one price observation per representative item, per outlet, and (often) per month
  - Pre-classified according to COICOP (Classification of Individual Consumption According to Purpose)
  - Price collectors*

- Alternative data
  - Web scraped data, API (pre-defined datasets), transaction data or other
  - Complete product universe – all items available (or purchased)
  - Not classified according to our purpose

    - Purchasing routines, shelf-placements or other product grouping relevant for the retailer, not CPI

Transaction data

- Detailed individual-level measurement of purchase
- Single transactions
- Price and quantity data generated at the time of purchase

Date	Time	Product ID	Product	Price	Quantity	Payment
2/10/2025	10:15	43165298	COCA COLA 1.5L	3.50	3	cash
2/10/2025	10:15	36987512	SODA 0.33L	2.00	1	cash
2/10/2025	13:14	40326897	MILK NO FAT 4 GALLON	2.00	1	credit card
2/10/2025	15:00	93568746	ORANGE JUICE, LOW ACID 1	4.00	2	credit card
2/10/2025	15:00	93568748	ORANGE JUICE, NO PULP, 1L	7.00	1	credit card
2/10/2025	16:01	43165298	COCA COLA 1.5L	3.95	1	cash
2/10/2025	16:10	36986791	GREEN APPLES, 1KG	4.50	0.5	cash

Scanner data/barcode data ++

- *Aggregated* transactions data
  - Total turnover and quantity sold for a given item by item code (GTIN, EAN or similar) for a given time period (often day or week)

Period	Product ID	Product	Turnover	Quantity	*Price**
20251039	43165298	COCA COLA 1.5L	206.40	60	*3.44*
20251039	93568746	ORANGE JUICE, LOW ACID 1	34.15	5	*6.83*
20251039	40326897	MILK NO FAT 4 GALLON	50.00	25	*2.00*
20251040	43165298	COCA COLA 1.5L	8.52	2	*4.26*
20251040	93568746	ORANGE JUICE, LOW ACID 1	90.15	15	*6.01*
20251040	40326897	MILK NO FAT 4 GALLON	231.20	68	*3.40*

Transaction data

- Superior data source!
  - Price and turnover information
  - Complete product universe and better data coverage
  - Actual transactions, including any point-of-sale discounts
- But..
  - Dynamic product universe
    - Product life cycles, life cycle pricing, technological development
  - Comparing like-with-like is less straight forward
    - Point-in-time replacements, quality adjustment, automation
  - Index formula, - weights or no-weights?

Photo by Michael Walter on Unsplash

Statistisk sentralbyrå
Statistics Norway

Two main approaches

- Static approach
  - Mimicking traditional data collection by collecting specified representative items by item ID
    - Utilize turnover information for selecting the relevant item variety
  - Easy to implement, better control of data, but limited use of the potential in the data

- Dynamic approach
  - Capturing the entire product universe
  - Monthly chaining to capture new items in a timely manner
  - Classification procedures, item definition and quality adjustments are more challenging
  - Can be complex in interpretation and code, however a better representation of the price movements experienced by households

Static approach

- No real change in traditional CPI methodology
  - Transaction data as a pure *data collection method*
  - Fixed sample, match GTINs over time, replacements are made

- More control over the actual items in the sample
  - ..but a limited sample!
  - Important to keep the basket relevant and replace items that are no longer relevant
    - Non-relevant or missing items must be replaced, and quality adjustment applied when appropriate

One GTIN (item ID) matched to each representative item, e.g.:

"*Mustard, 500g*"

Item ID	Item text
7392031000016	JOHNNY'S SENNEP SØTSTERK 500G
7039010081047	IDUN SENNEP GROV 275G
3036810201372	MAILLE DIJON SENNEP 380G
7039010081061	IDUN SENNEP STERK 250G
7052860000549	BERGBYS SVARTPEPPER SENNEP 250G
7052860000433	BERGBYS SENNEP SØT 250G
7039010082426	**IDUN BODSENNEP 490G**
7023256006957	CHATEL DIJONSENNEP LYS 370G
7052860000570	BERGBYS GUL AMERIKANSK SENNEP 250G
3230140002719	FALLOT SENNEP DIJON ESTRA.210G
7052860000518	BERGBYS CURRY-MANGO SENNEP 250G
711381000441	SENNEP CHAMPAGNE 226G STONEWALL
7035620032721	SENNEP GROV 180G JACOBS UTVALGTE
7350012370107	SØT&STARK SENNEP 150G LISSELLAS
7071281197461	VIKING RAGGEN 746 SENNEP 150G
7310700660323	SENNEP 290G SLOTTS
5711111041220	JULESENNEP 200G CERVERA
5900084273053	SENNEP M/HONNING 185G KAMIS
⋮	⋮

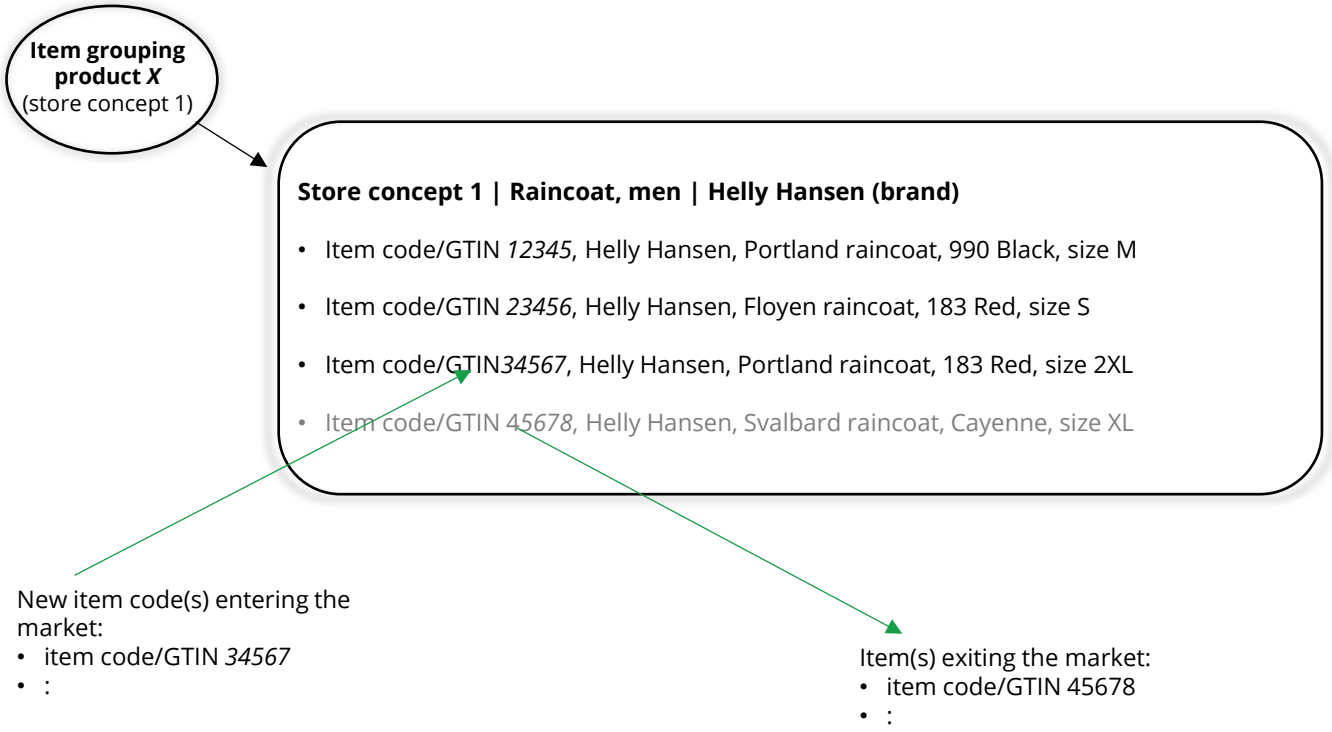
Dynamic product universe, dynamic approach, but..

- Product lifecycles
  - Stability in item codes over time
    - Food and non-alcoholic beverages, pharmaceutical products, petrol..
  - Short-lived item codes
    - Clothing, shoes, consumer electronics..
- Life cycle pricing
  - The item enters the market at a starting price, mid-life prices varies, and then the item often leave the market at a discount
    - Important to link new items to the outgoing item or else the index will have a downward bias
- Technological development
  - Consumer electronics are often short-lived, prone to life cycle pricing, and new items are of different quality....

Item definition

- Stable product codes, no life cycle pricing
  - An item is defined by a unique item ID (GTIN, EAN, retailer specific code or other)
  - Matched model approach of items over time at item code level
  - *Relaunches* must be linked to the outgoing item code (capturing shrinkflation)
- Short-lived products, life cycle pricing, no technical development
  - Item varieties change, but the product universe is stable (i.e. t-shirts, pants)
  - Matched model approach of *groupings* of individual items of similar type and quality over time (unit value)
  - Items enter and exit the grouping throughout, no need for manual linking
- Short-lived products, life cycle pricing, rapid technical development
  - … hedonic methods? Homogenous product groups? Other?

Homogenous product groups, example



Statistisk sentralbyrå
Statistics Norway

Index formula – weights or no-weights

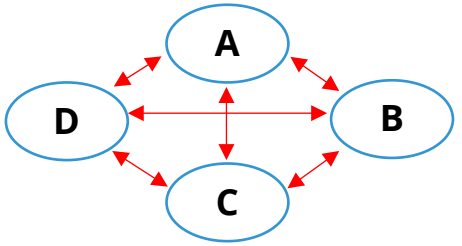
- Using explicit weights combined with monthly chaining can introduce a bias into the index (so-called *chain drift*)
  - Month t — Pizza on sale, quantity increases → price decrease heavily weighted
  - Month t+1 — Pizza returns to normal price, but quantity is less because freezer is full → the index in *month t+1* does not return to the same level as *month t-1*
  - Month t+2 — Pizza at normal price and quantity is back to normal → no price change from *month t+1* to *month t+2*, the index is stuck at lower level..

- Solution: unweighted indices using cut-off or other filters to ensure a relevant sample of goods..

..Or: **multilateral methods**

Multilateral methods

- Utilize the full potential in the data
  - Dynamic product universe – capturing new items at a timely manner
  - Including weights (expenditure shares) at detailed levels without creating bias
- Traditionally used to compare price levels across countries
  - Direct price comparisons between country A and C should give same results as an indirect comparison of country A with country C via country B



- Adjusted to comparison over time: $Index_{1,3} = Index_{1,2}\ x\ Index_{2,3}$

- In short: we use data from multiple periods, not just two (bi) periods

Multilateral methods cont.

- Usually based on superlative indices using expenditure data from two periods, such as bilateral Törnqvist or Fisher indices. However, can also be used for unweighted indices such as Jevons
- A time window is used to avoid revisions
  - Superlative indices within a defined time window where each period is used as a base period
  - Geometric mean of the Törnqvist (or Fisher) indices

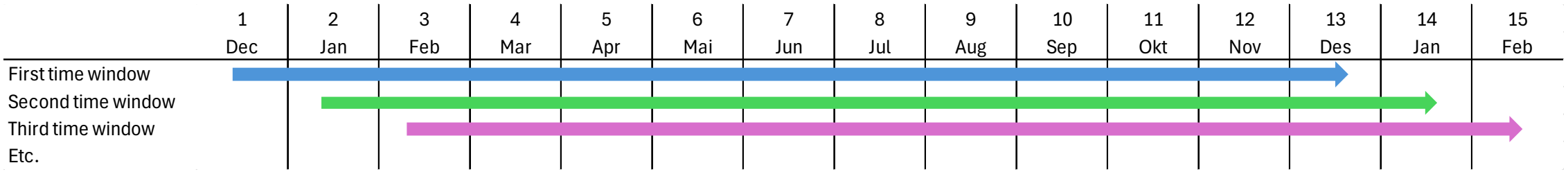
$$P_T^{0,t} = \prod_{i \in S} \left(\frac{p_i^t}{p_i^0}\right)^{\left(\frac{s_i^0 + s_i^t}{2}\right)}$$

$$P^{1,2} = \left(\frac{P^{0,2}}{P^{0,1}} x \frac{P^{1,2}}{P^{1,1}} x \frac{P^{2,2}}{P^{2,1}}\right)^{1/3}$$

Time window

- The length of the window may vary depending on product group and data availability
  - Capture seasonal items, but will need stability in available items the longer the period
  - The longer the window, the more back data you might need when the method is being introduced

Fixed length rolling window

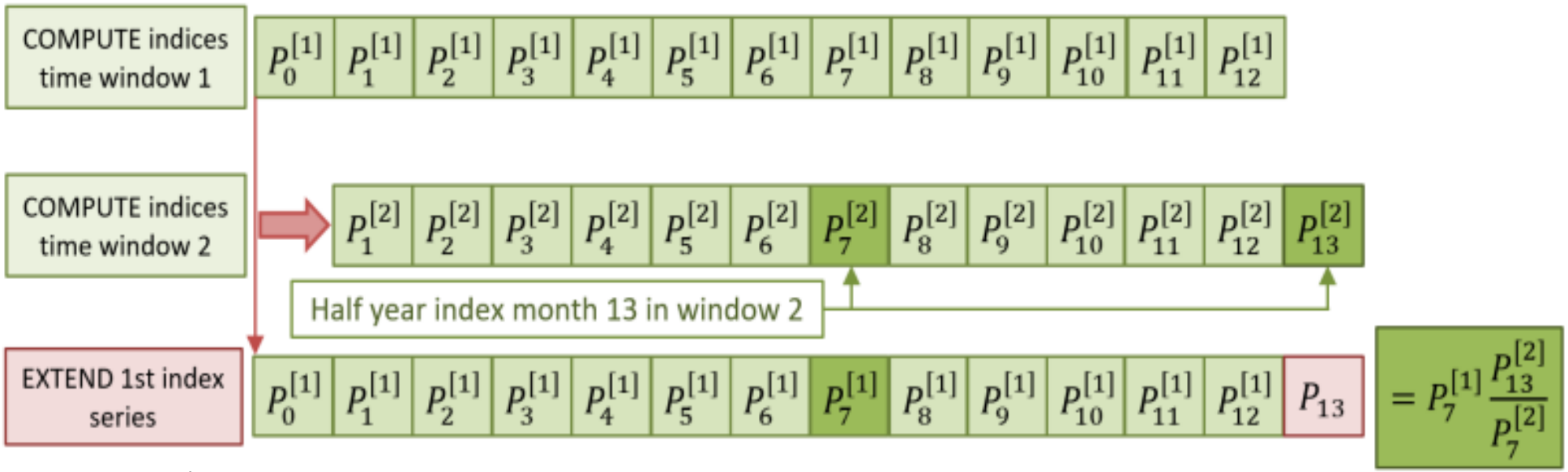


Norway: GEKS Törnqvist (CCDI), halfsplice

- **2021**: Food and non-alcoholic beverages, **2022**: fuels and non-food from supermarkets, kiosks and fuel stations
- Törnqvist price indices within a time window of 25-months, each period used as a price reference period
- Half splice
  - Geometric mean of all Törnqvist indices for current period compared to the same period previous year = 12-month rate of change

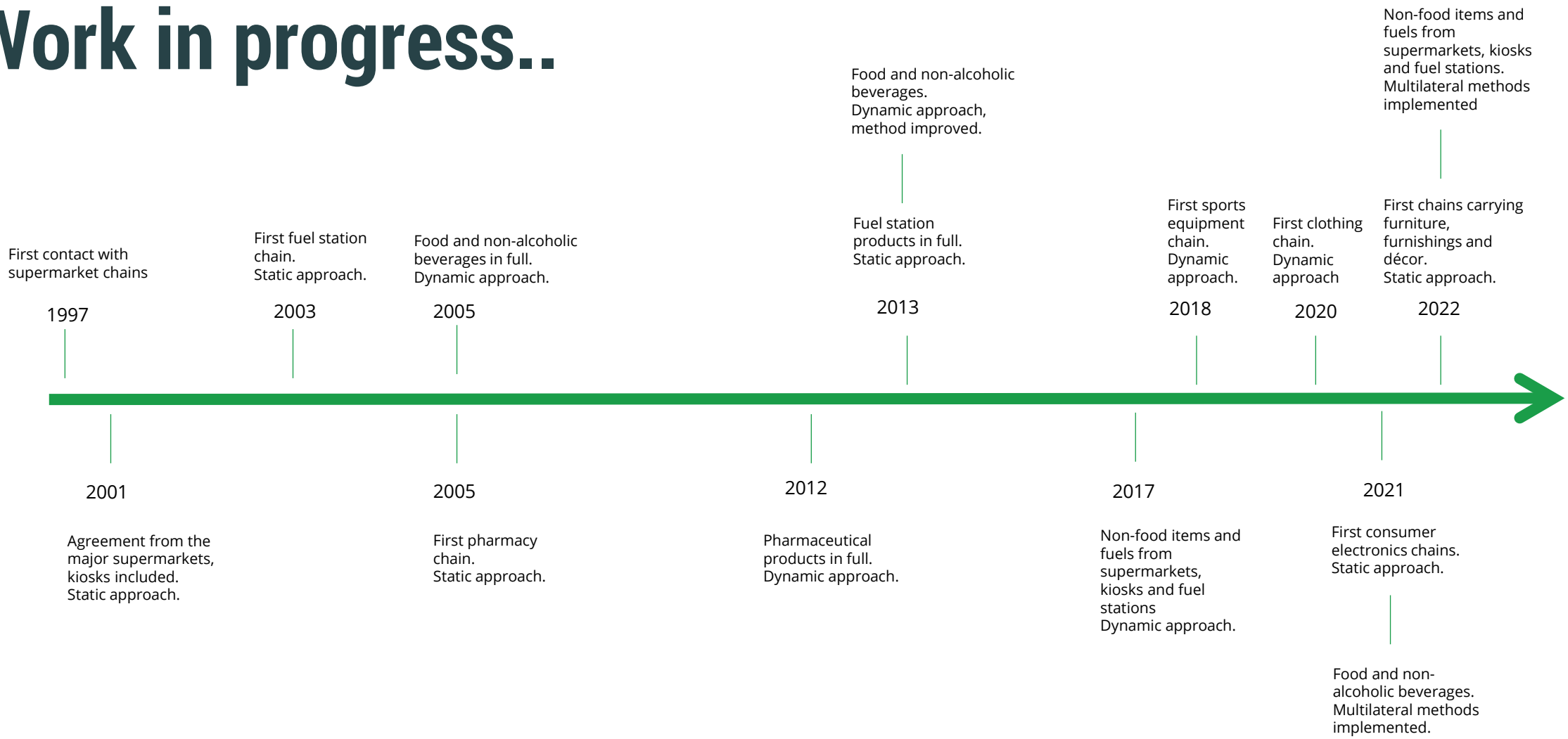
$$P_{GEKS}^{Oct23,Oct24} = \left(\frac{P^{Oct22,Oct24}}{P^{Oct22,Oct23}} x \frac{P^{Nov22,Oct24}}{P^{Nov22,Oct23}} x \frac{P^{Dec22,Oct24}}{P^{Dec22,Oct23}} \dots x \frac{P^{Oct24,Oct24}}{P^{Oct24,Oct23}}\right)^{1/25}$$

  - The 12-month rate of change is chained to published figures in period t-12



Source: Antonio Chessa, NL

Work in progress..



Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Thank you for your attention

Statistisk sentralbyrå
Statistics Norway