Online Price Index with Product Replacement: The Closest Match Approach

Manuel I. Bertolotto

PriceStats & Universidad de San Andrés

May 9th, 2019

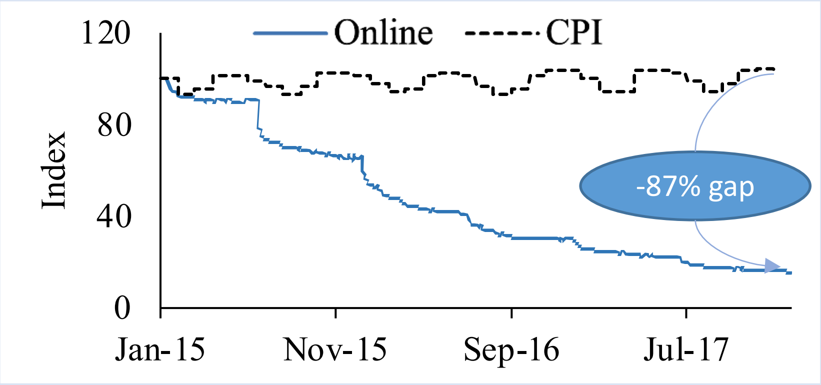
Outline

1. Key problem: inflation rate measures using large datasets show abnormal trends
2. Factors contributing to this problem
3. Why don’t consumer price indices with traditional datasets suffer from this problem?
4. Possible solution to this problem: new approach to calculate a price index using large datasets – the Closest Match
5. Conclusions

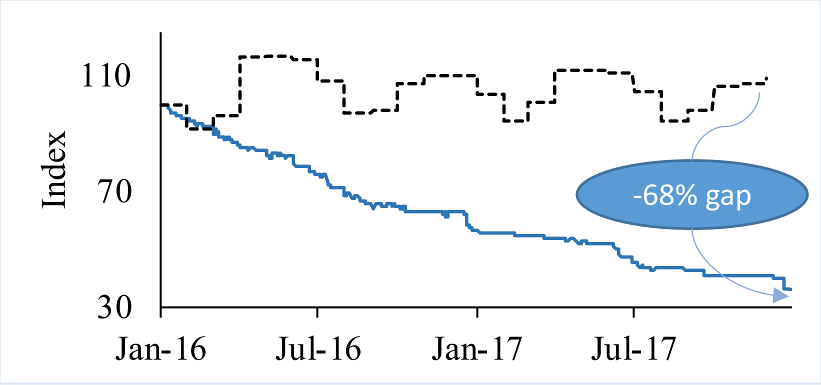
Current methodologies to calculate a consumer price index (CPI) with online prices have shown an abnormal downward trend

Examples: Apparel CPI, non-seasonally adjusted

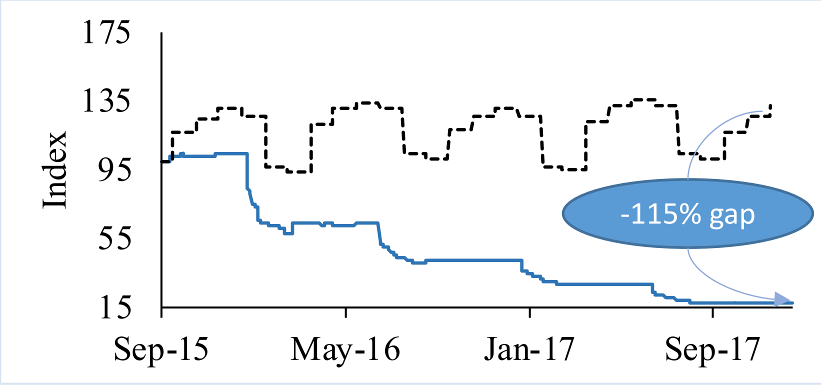
Germany



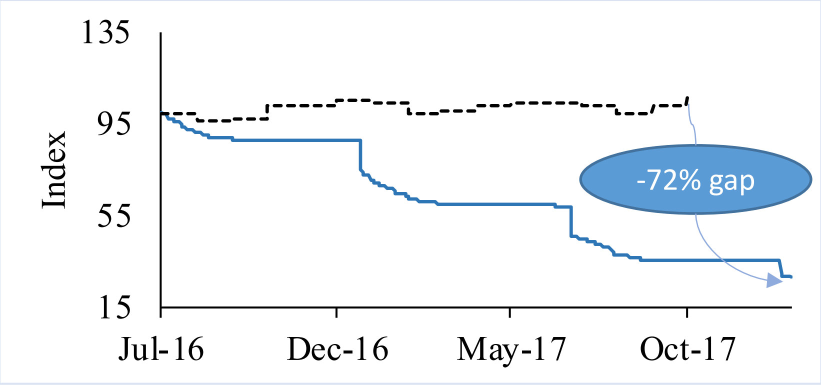
Netherlands



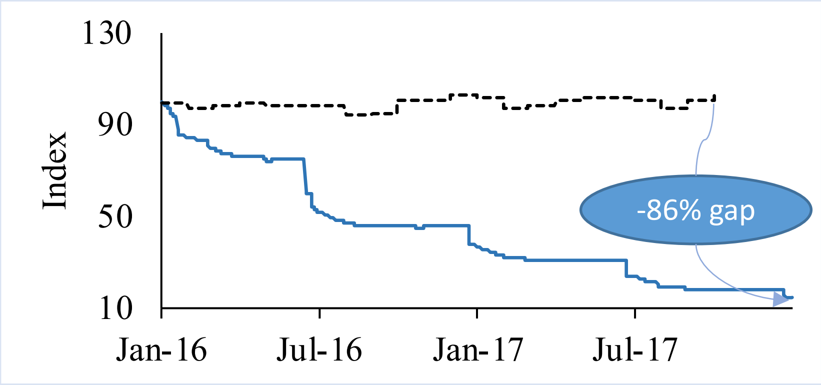
Spain



UK



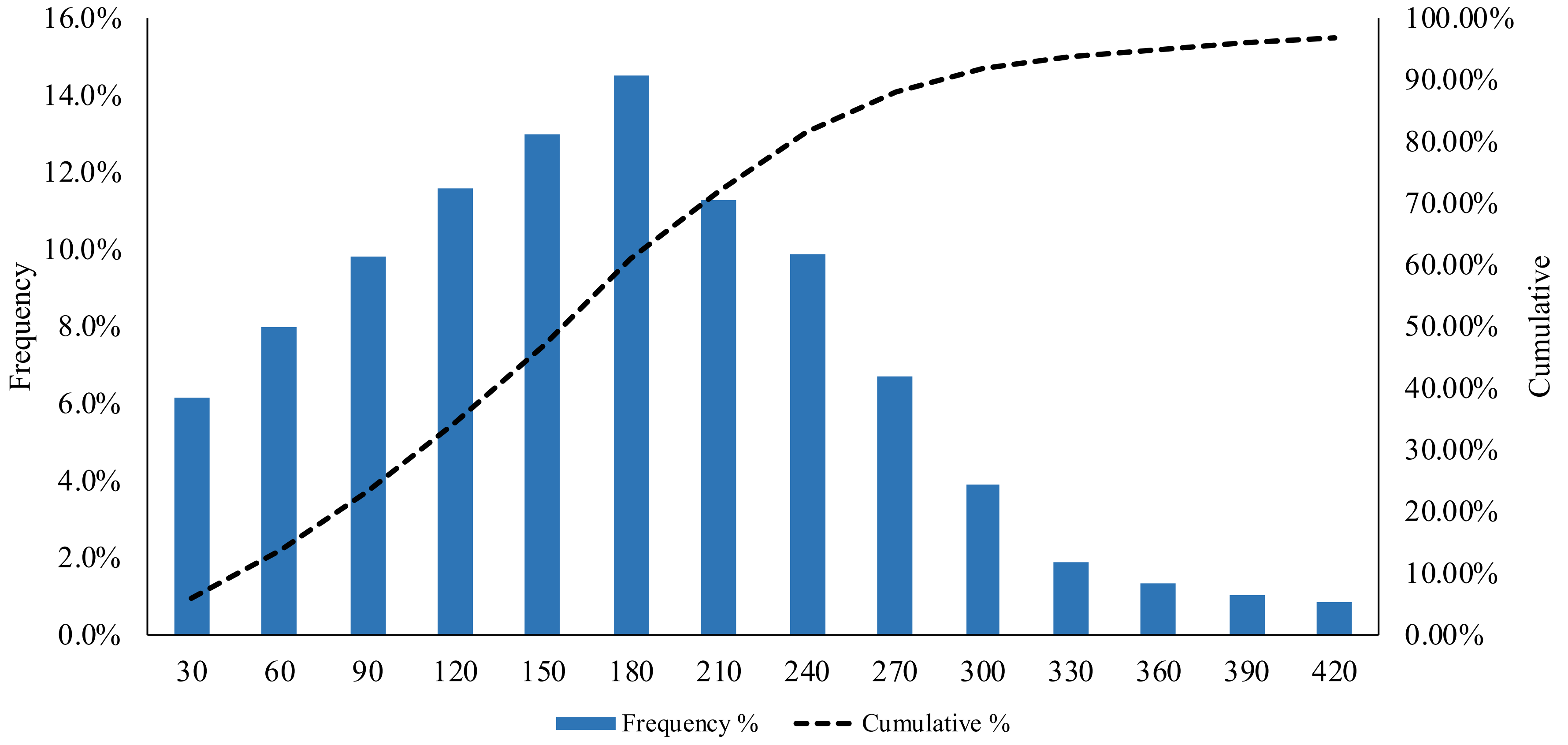
USA



Source: National Statistical Offices & PriceStats

Three factors explain the downward trend of the online indices: First, retailers replace more than 90% of their products every year

Distribution of the lifespans of products

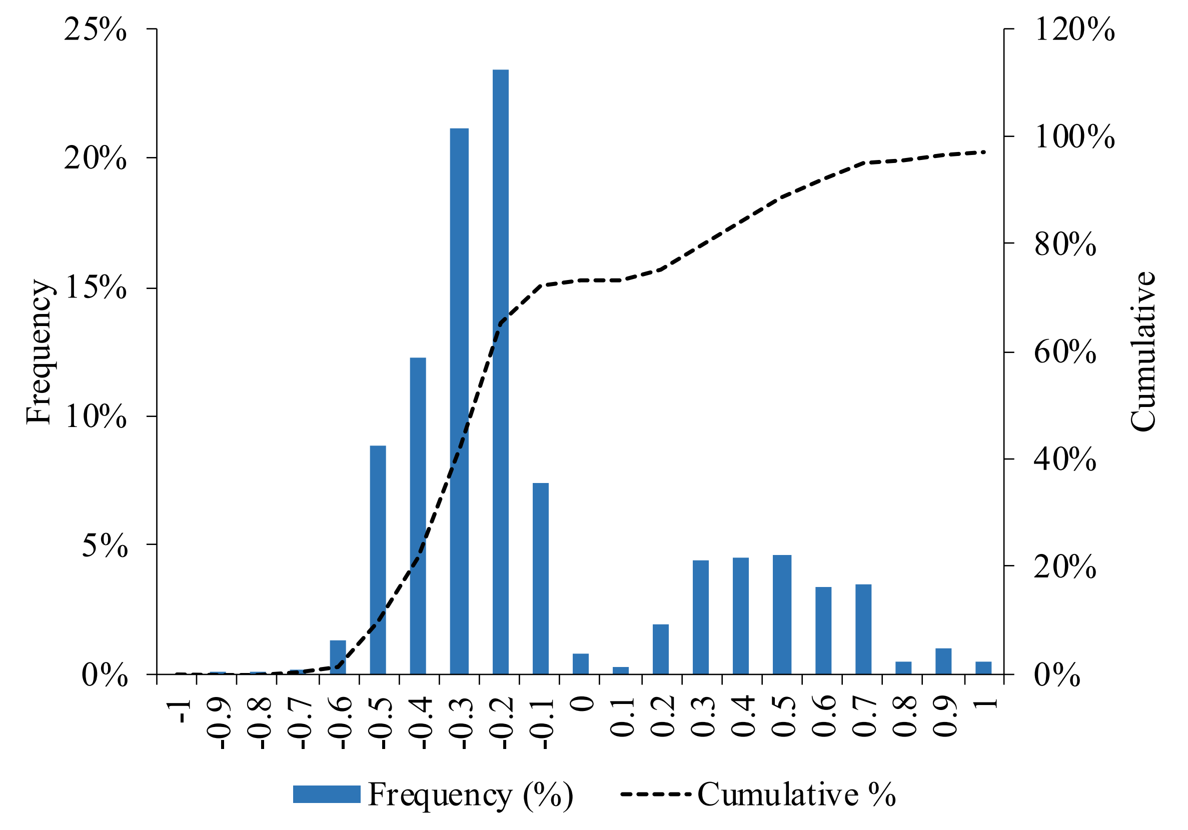


Distribution facts

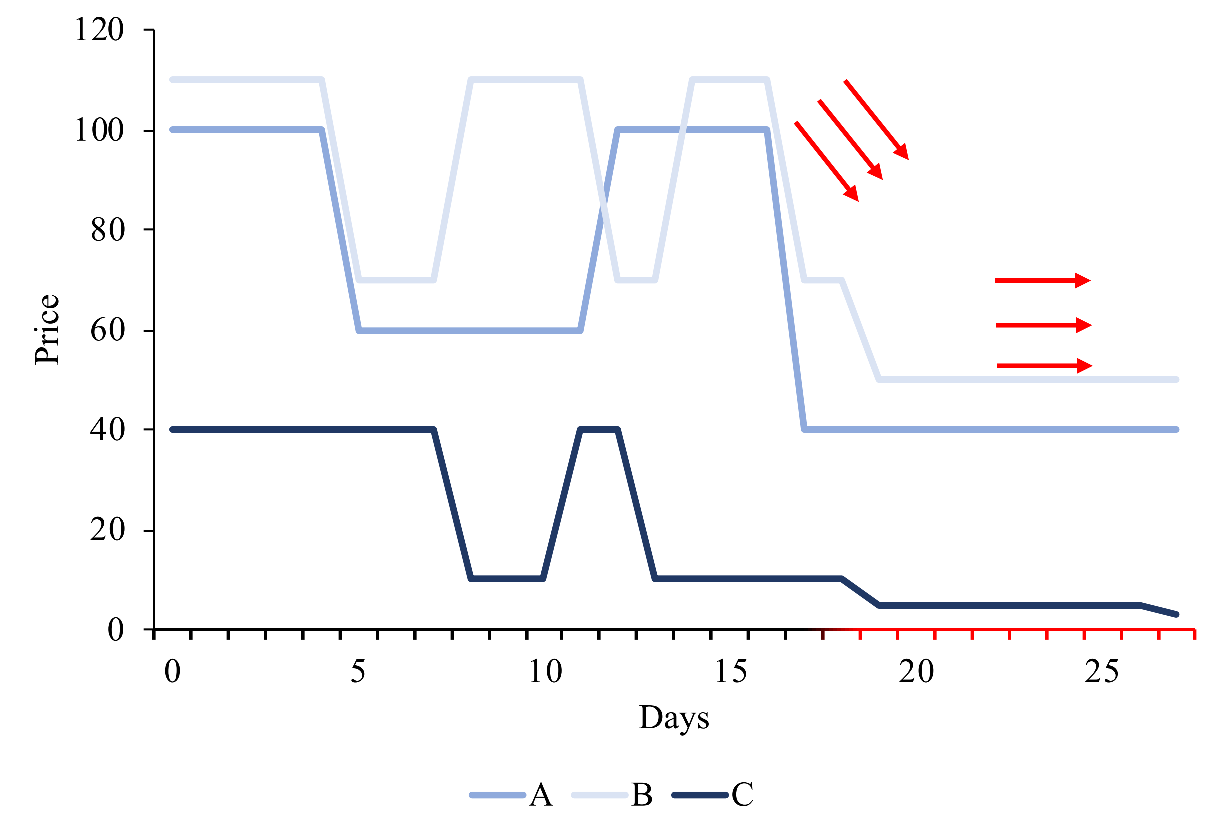
- 50% of products last 5 months or less
- 75% of the products last 7 months or less

Second, around 75% of the price changes are negative

Distribution of size of price changes

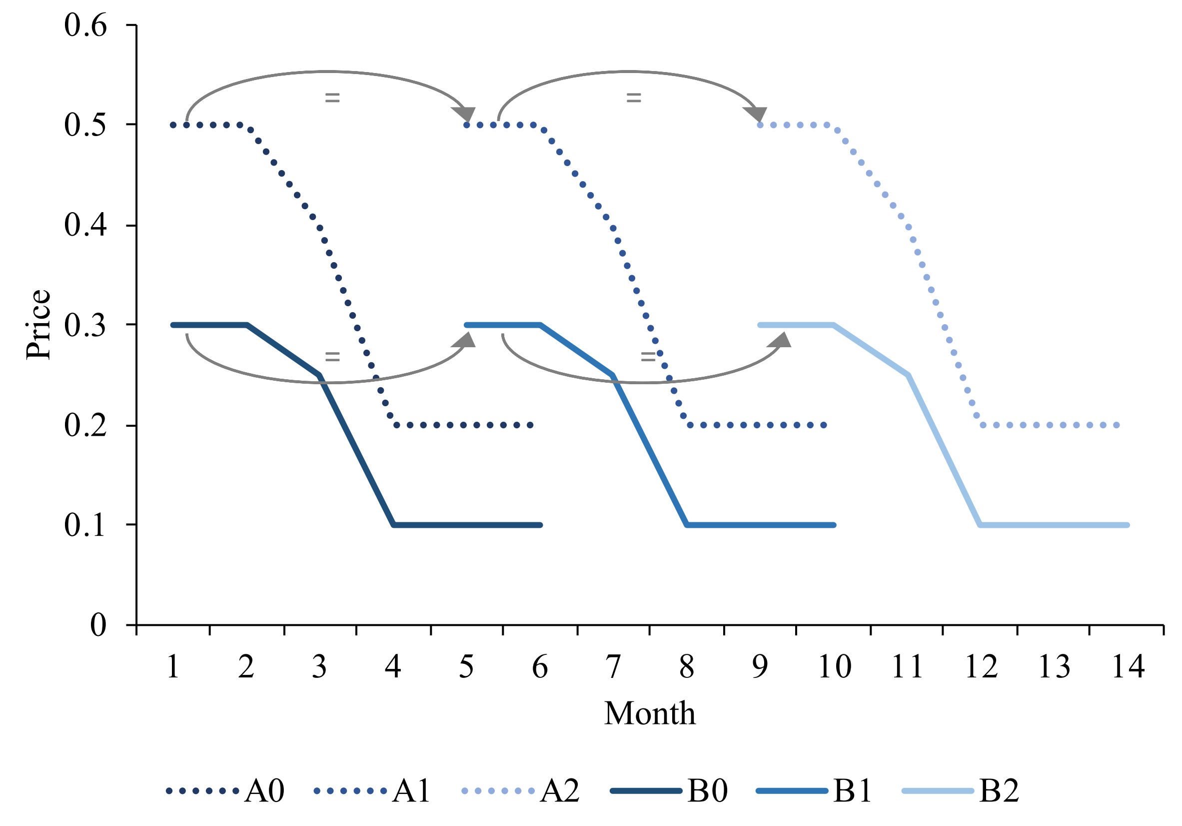


Typical sale pattern in the apparel sector

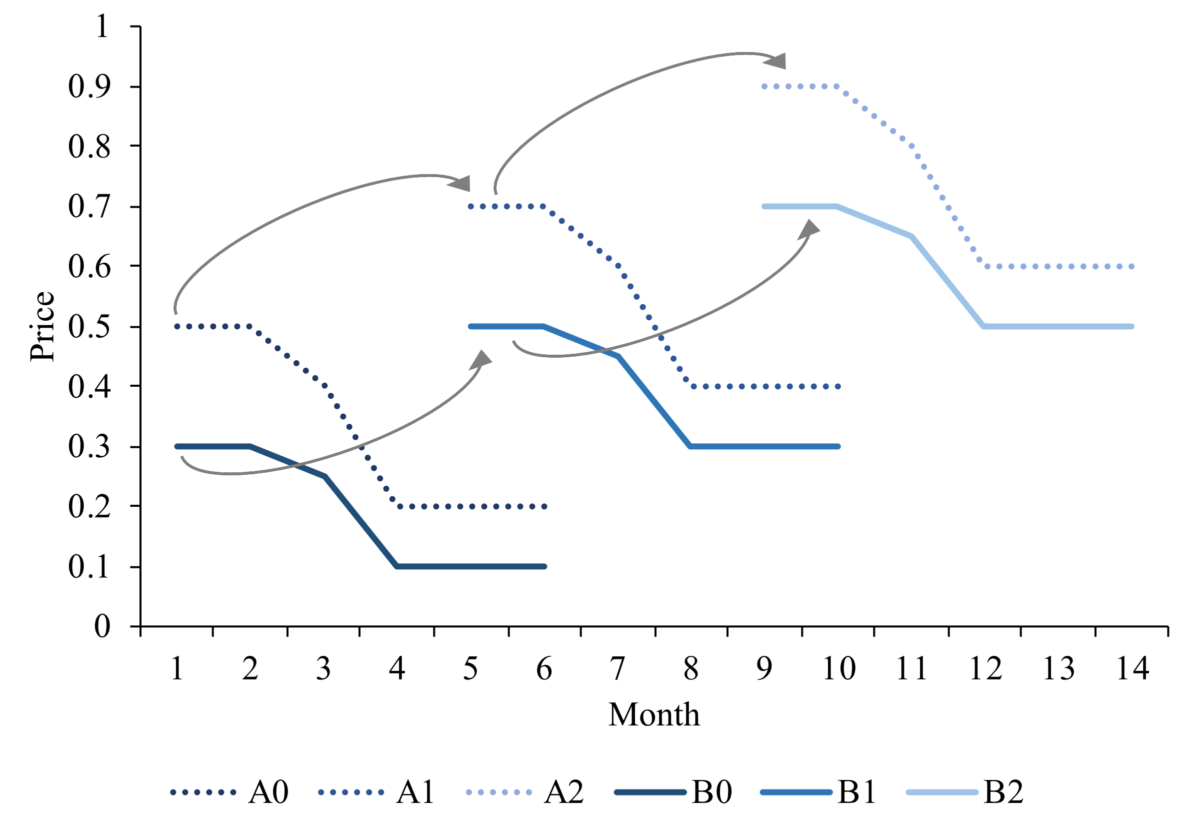


Third, products are usually introduced into the market at a full price and discontinued at a clearance price

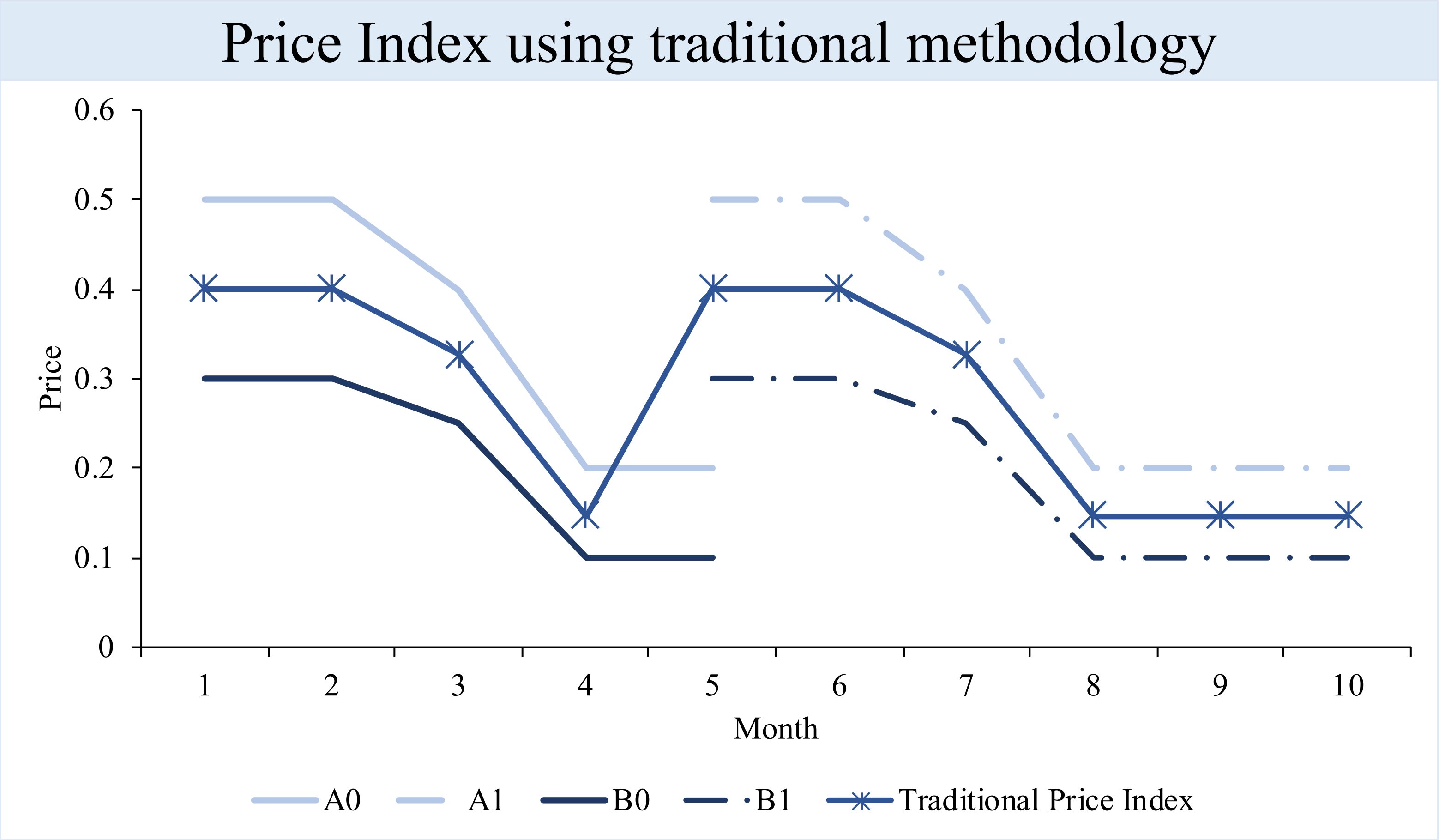
Model cycle with no inflation



Model cycle with inflation



These factors do not affect the traditional CPI due to the collection methodology of national statistical offices

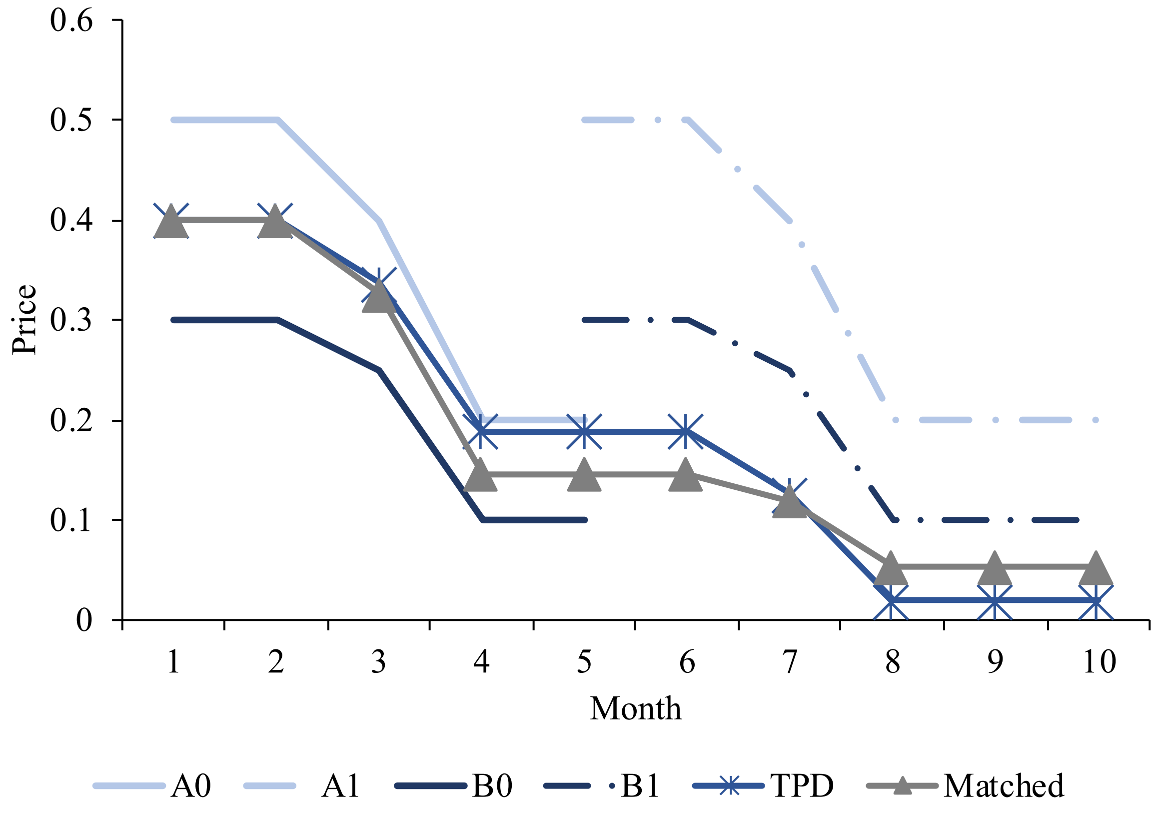


- The accumulated inflation rate of the CPI is:

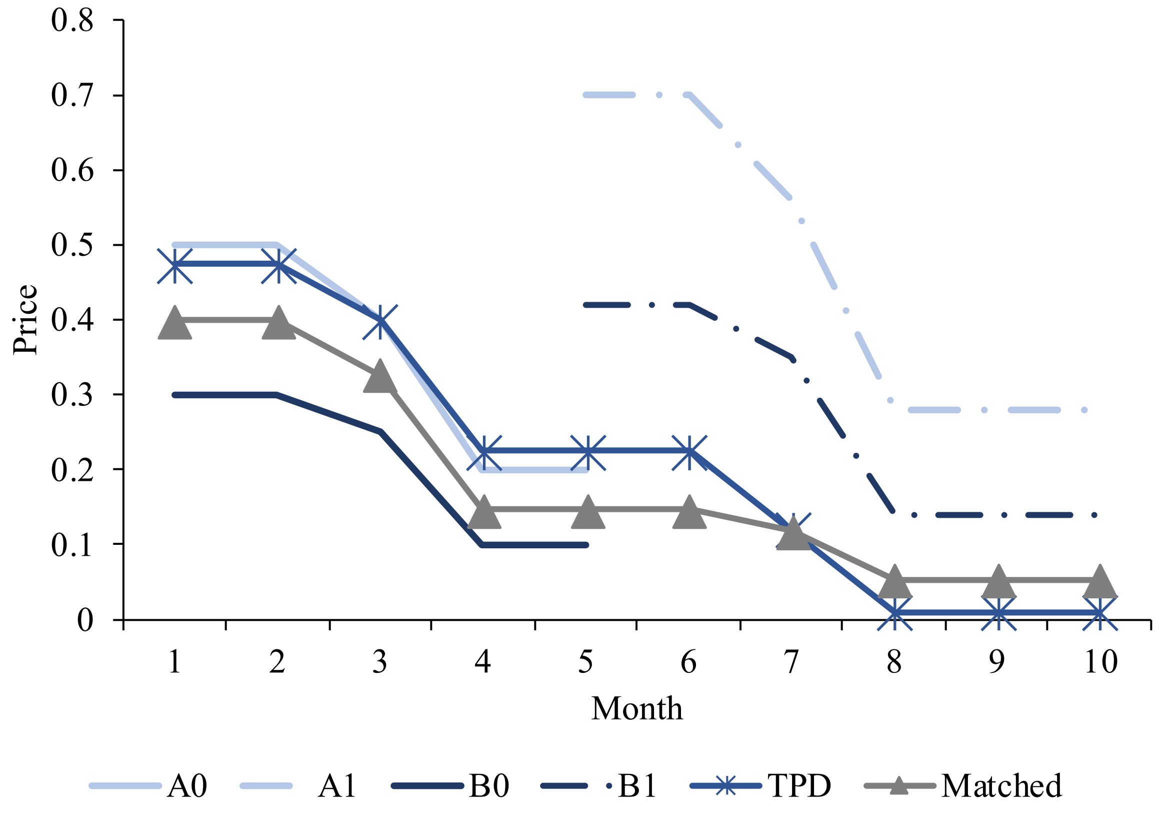
$$\Delta P^{JV} = \left(\frac{p_{A1,10}}{p_{A0,1}} \frac{p_{B1,10}}{p_{B0,1}}\right)^{\frac{1}{2}}$$

However, online methodologies cannot identify qualitatively similar goods. Instead, both the old and new models of a good are automatically assumed to be different products.

Price Index - 0 % inflation rate



Price Index - > 0% inflation rate



TPD stands for Time Product Dummy

The Time-Product Dummy (TPD) method exemplifies the identification issue

Model equation	$\ln(p_{i,t}) = \alpha + \sum_{t=2}^{10} \delta_t D_{i,t} + \sum_{i=1}^{3} \gamma_i D_i + \varepsilon_{i,t}$
Accumulated inflation	$\Delta P^{TPD} = \underbrace{\left(\frac{p_{A1,10}}{p_{A0,1}} \frac{p_{B1,10}}{p_{B0,1}}\right)^{\frac{1}{2}}}_{\Delta P^{JV}} \underbrace{\left(\frac{p_{A0,5}}{p_{A1,5}} \frac{p_{B0,5}}{p_{B1,5}}\right)^{\frac{1}{2}}}_{< 1}$

Identifying equal-quality products eliminates the abnormal downward trends of the online price indices

What happens when we identify equal-quality products using the TPD model?

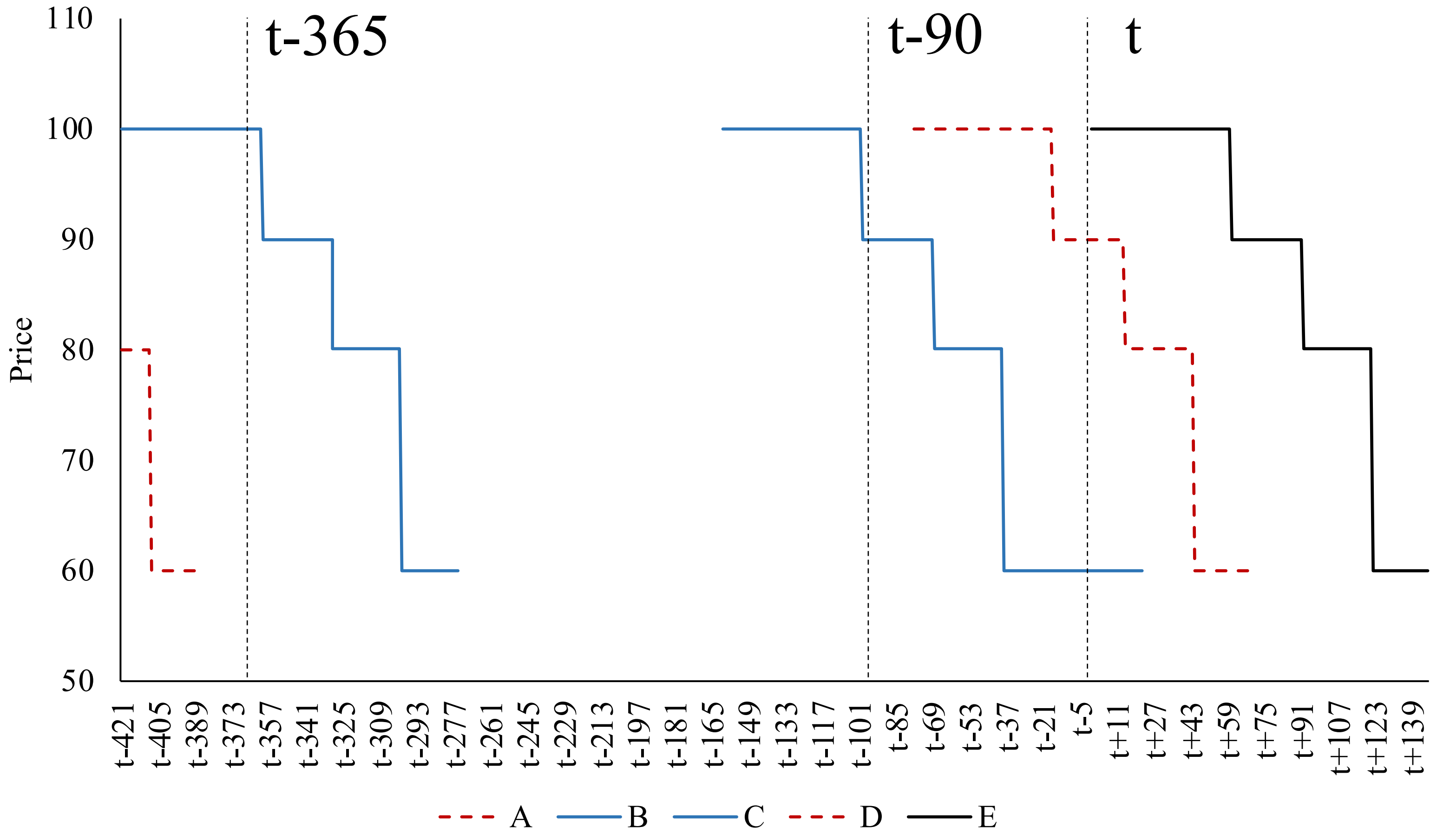
Model **and quality restriction**	$$\ln(p_{i,t}) = \alpha + \sum_{t=2}^{10} \delta_t D_{i,t} + \sum_{i=1}^{3} \gamma_i D_i + \varepsilon_{i,t}$$ $$\begin{cases} \gamma_{A0} = \gamma_{A1} \\ \gamma_{B0} = \gamma_{B1} \end{cases}$$
Accumulated inflation rate	$$\Delta P^{TPD-C} = \Delta P^{JV} = \left(\frac{p_{A1,10}}{p_{A0,1}} \frac{p_{B1,10}}{p_{B0,1}} \right)^{\frac{1}{2}}$$

The Closest-Match method searches for a comparable item every time a new product enters the market. The method is scalable and automated

1 Filter	• The objective is to identify products from the previous season • Reduce the number of computations required in the next step in the score calculation
2 Score	• Calculates a score for each feasible close-alternative good • The product with the highest score is considered the closest match, as long as that score is higher than a pre-defined threshold

Three rules define the filter

Potential replacements



Rules

1. The start date of the replaced item is earlier or on $t - 90$.
2. The end date of the replaced item is at most $t - 365$ days old.
3. The replaced item has been available for at least ten days in the data.

The score is higher when the two product descriptions are more similar

$$S(q,d) = r(q,d) * \sum_{w=1}^{N} idf(w) \cdot fln(w,d)$$

q, d, w	Newly introduced item, feasible close-alternative good, word, respectively
$r(q,d)$	Relevance: Number of common words out of the total number of words in the newly introduced item (q)
$idf(w)$	Inverse description frequency of word *w*
$fln(w,d)$	Inverse of the number of words in a product description of the close-alternative good

The score threshold is defined ex-ante, based on a random sample of products

Process

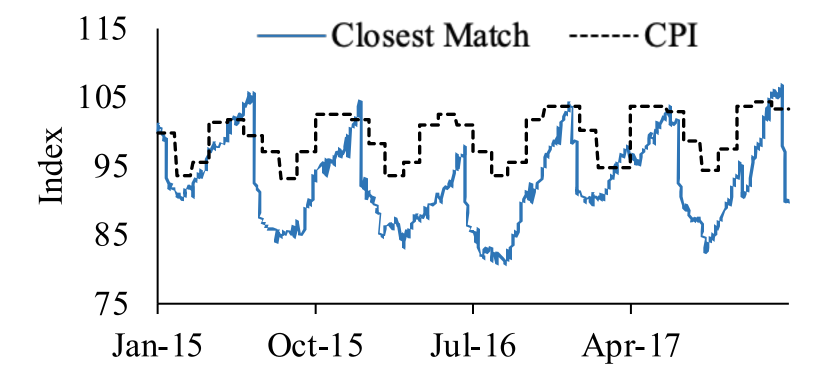
1. Select a random set of items and find their closest match
2. Sort the list of products by the score
3. Identify a threshold where products with a higher score than this threshold are of similar quality, and products with a lower score are significantly different

Example

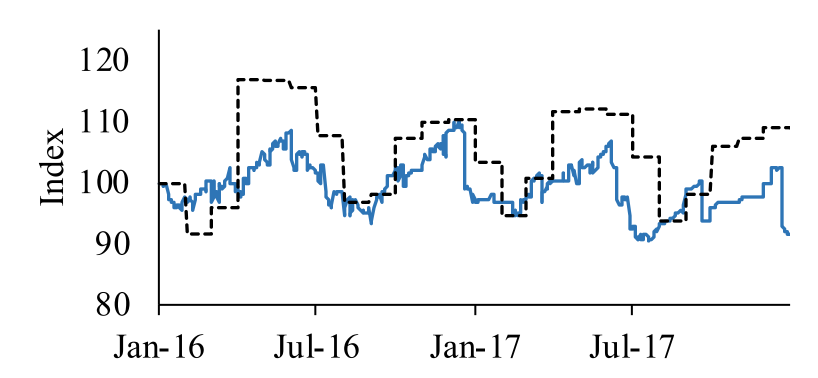
New product	Closest match	Score
• V-neck Blouse, dark blue	• V-neck Blouse, dark blue	13
• Off-the-shoulder Blouse, cotton	• Off-the-shoulder Blouse, cotton	13
• <u>Blue</u> shirt, 100% cotton	• <u>Red</u> shirt, 100% cotton	9
• <u>Patterned Viscose</u> Blouse	• Blouse <u>with Butterfly Sleeves</u>	2.5

The Closest-Match Indices are remarkably similar to the traditional CPI

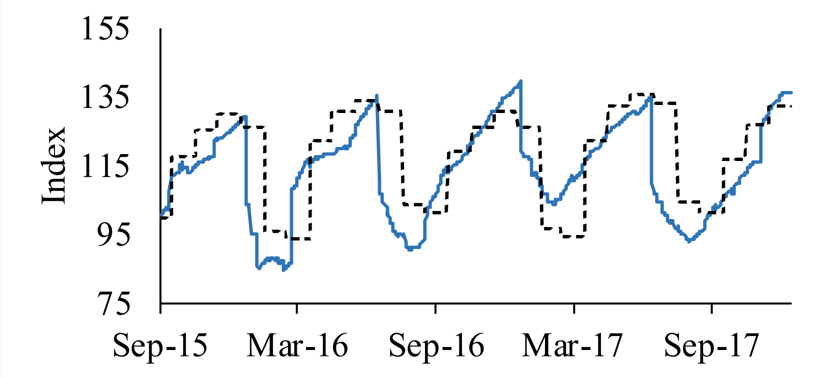
Germany



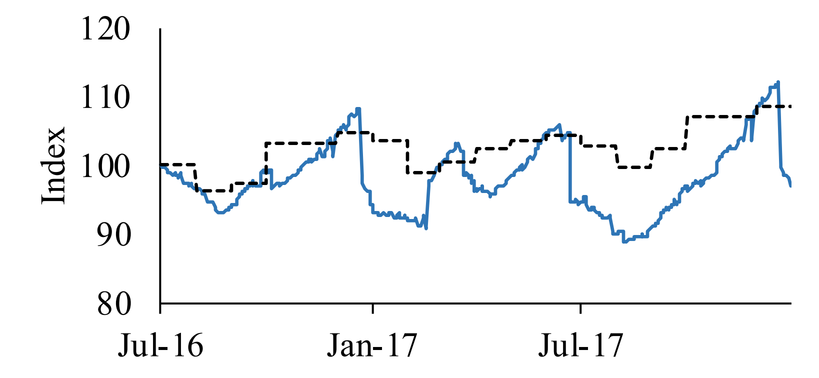
Netherlands



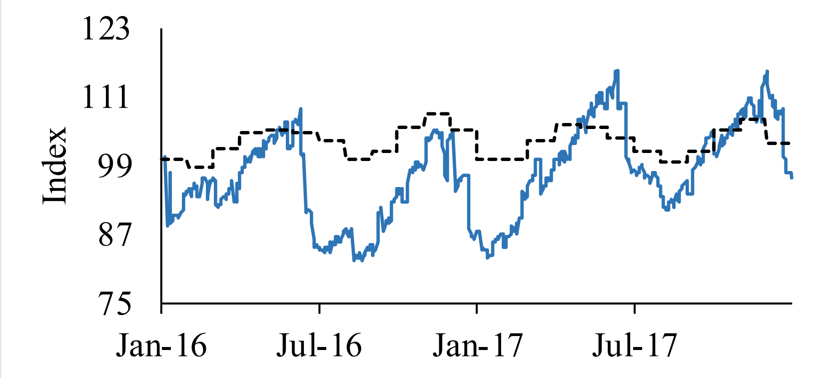
Spain



UK



USA

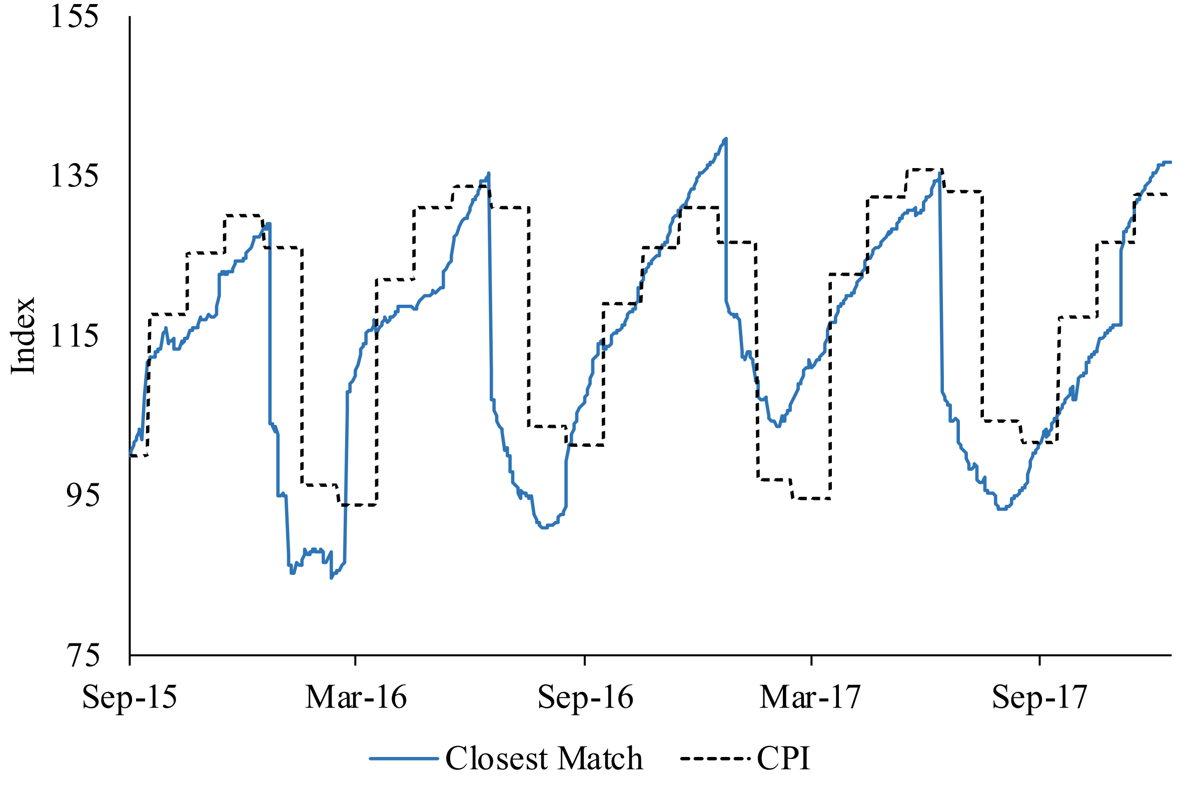


Source: National Statistical Offices & PriceStats

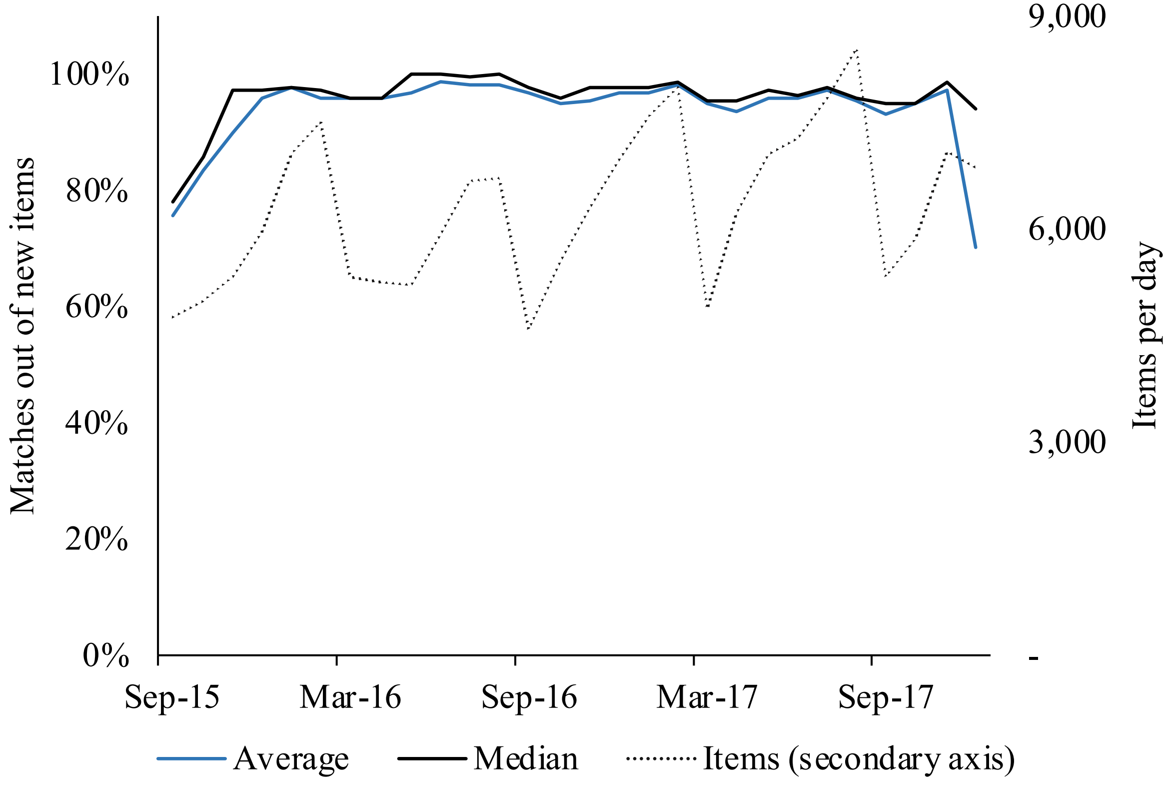
The online price index is not affected by the typical volatility of the number of items included

Example - Spain

CPI versus Closest Match Index



Matches Out of New Items and Items per Day

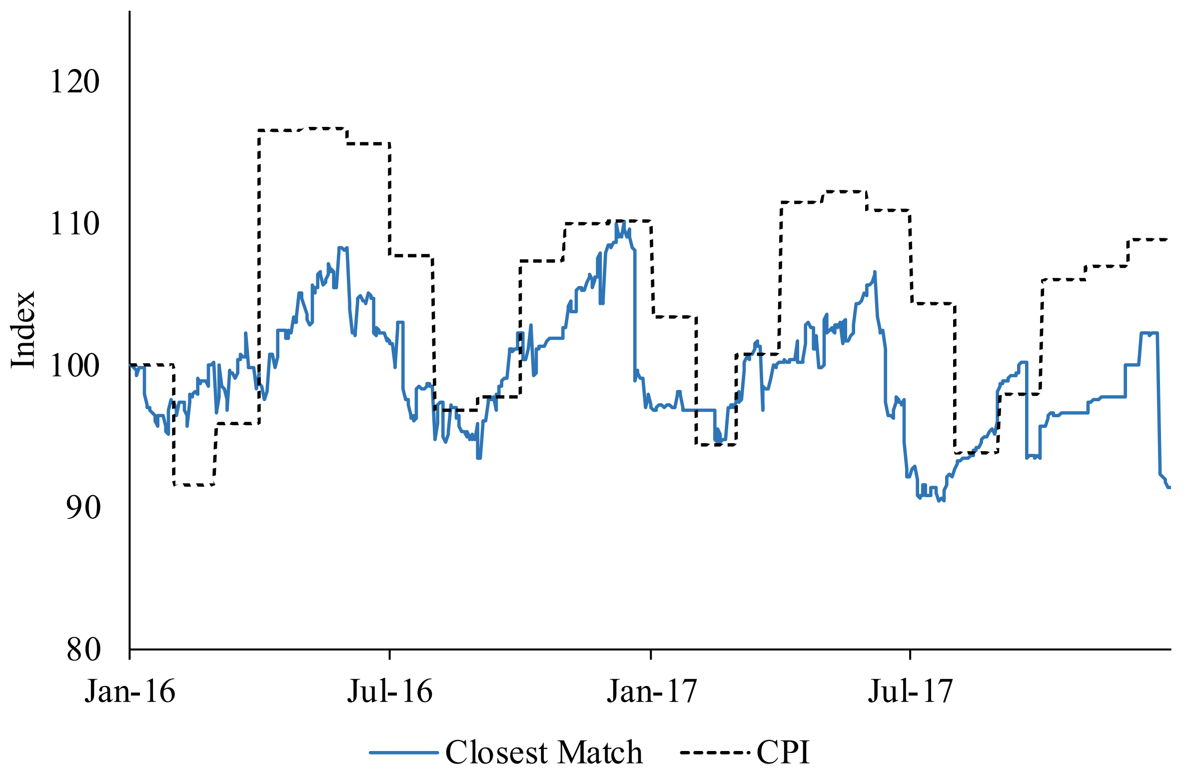


Source: Instituto Nacional de Estadística (INE) & PriceStats

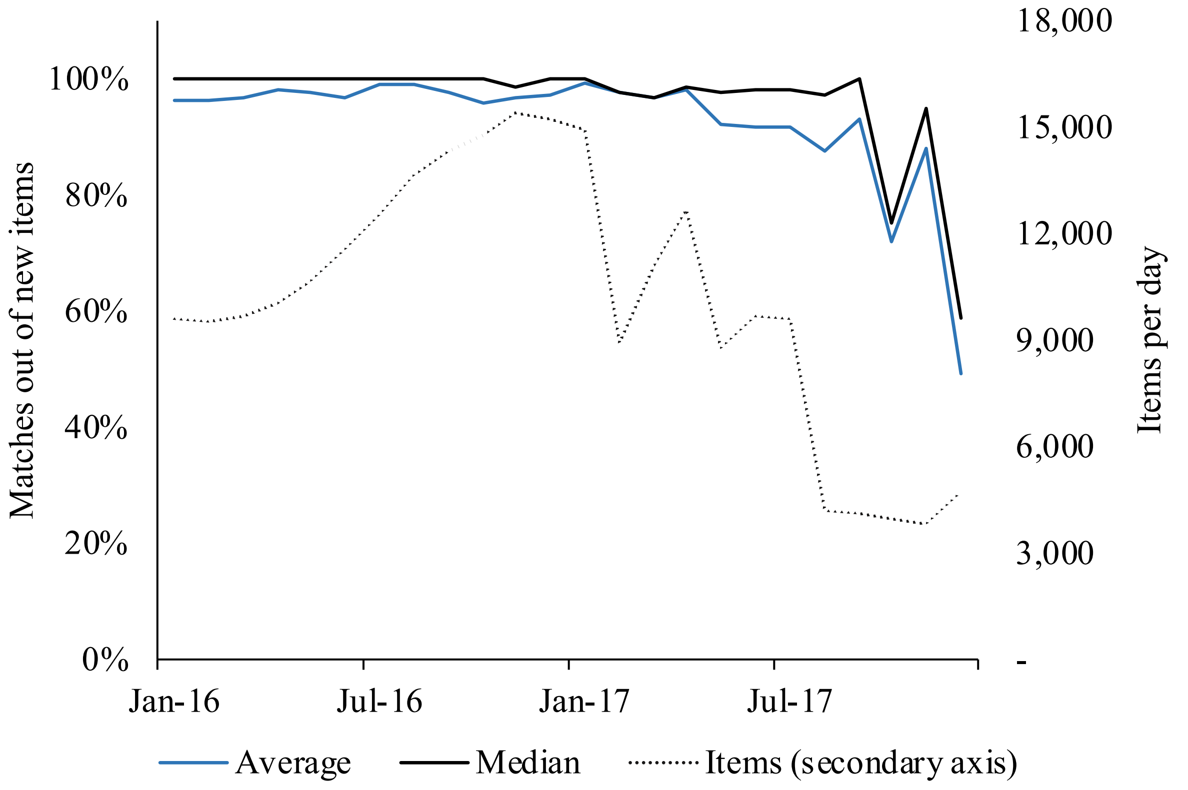
A persistent and large drop of the number of items negatively impacts the performance of the method

Example - Netherlands

CPI versus Closest Match Index



Matches Out of New Items and Items per Day



Source: Centraal Bureau voor de Statistiek & PriceStats

Conclusions

- I suggest a possible solution to the product turnover problem: the Closest Match approach
- Instead of looking for a replacement when an item is discontinued, the closest-match approach searches for a comparable item every time a new product enters the market
- The methodology is robust to the typical item-count volatility of online data sources
- The method is scalable so that price indices can be calculated with thousands of products without manual intervention
- The closest-match indices show remarkably similar inflation trends to the traditional CPI