Online Price Index with Product Replacement: The Closest-Match Approach

Manuel I. Bertolotto

Universidad de San Andrés and PriceStats

mbertolotto@udesa.edu.ar

December 2018

Abstract

Current methodologies to calculate a consumer price index (CPI) with online prices have shown an abnormal downward trend. This paper introduces a new methodology that avoids this problem by effectively mitigating the effects of product turnover. The method mimics the decision-making process of the specialist that reviews forced-replacement items at the statistical offices, yet the method is scalable so that the price index can be calculated with thousands of products without manual intervention. The method reflects a change in paradigm for how old and new varieties of products are linked. While the traditional approach looks for a replacement when an item is discontinued, the closest-match approach searches for a replaced item every time a new product enters the market.

The price index presented in this study is remarkably similar to the traditional CPI for every country in the sample, namely Germany, the Netherlands, Spain, the United Kingdom, and the United States.

JEL: C43, C82, E31.

Keywords: online price index, product introduction, quality change, scraped prices.

1. Introduction

The availability of online prices has led to a growing number of papers studying how this new source of data can be used to measure the Consumer Price Index (CPI). Online price indices have many advantages but deciding how to deal with product turnover is a key remaining challenge to their implementation. In particular, when a good disappears from a store, data collectors at the national statistical offices (NSO) look for a comparable replacement to continue measuring the good's price trend. Online methodologies do not account for such product turnover, so the price indices for some CPI categories have abnormal trends (see De Haan & Hendriks (2013)).

This paper presents a method that effectively deals with product turnover. The problem that needs to be solved is that existing online methodologies are not able to identify qualitatively similar goods. Instead, both the old and new models of a good are automatically assumed to be different products, causing a downward bias in the price index because new models tend to be sold at higher prices than the older models. While this issue exists in most of the CPI categories, the apparel sector prominently exemplifies its consequences. Items in this sector are usually introduced into the market at a full price and discontinued at a clearance price. The online indices capture the visible price decreases during the lifetime of each product but fail to identify the implicit price increases that consumers face when new models are introduced into the market. As shown in Section 2, around 60 percent of products have a lifecycle shorter than six months. Given such short life expectancies, it is likely that the new models are almost identical to their predecessors and should be considered replacements.

To deal with product turnover, the paper presents a method that automatically matches each newly introduced item with its closest alternative good from the existing pool of products. It then

compares the first price of this new item against the last price of its closest alternative good and records this as the new item's first price change. This method, which I label "closest-match", mimics the decision-making process of the CPI experts who currently review item replacements at the statistical office.

My method identifies replacement items automatically, so the number of subjective decisions made by analysts calculating a CPI is minimized. Additionally, this method is scalable, so it is possible to increase the number of items in the price index without significantly increasing the number of agents dedicated to overseeing item replacements. Moreover, as suggested in Section 3, the method complements existing methodologies to calculate the online price indices, such as the fixed-effects model proposed by Krsinich (2016). Contrary to previous results, the online price index presented in this study yields similar results to the traditional CPI.

This work is related to a strand of literature using alternative sources of information to calculate consumer price indices. Cavallo & Rigobon (2016) show that an online price index using the overlapping-quality methodology presents similar inflation trends to the traditional CPI. Under this approach, forced replacements are unnecessary because a price gap at the time of introduction of the new variety reflects a quality difference. Krsinich (2016) proposes a quality-adjusted online price index estimated by a fixed-effects model, which presumably avoids forced replacements as well. When a new product is introduced into the index, the fixed-effect model decides how much of the price gap comes from quality differences and how much reflects inflation. Goolsbee & Klenow (2018) calculate an online price index that suggests that the CPI might overestimate inflation by underweighting the role of new products. I build on this literature to suggest a complementary method to calculate online price indices that mitigates a product-turnover bias.

There are also studies using scanner data to construct price indices, such as De Haan & Van Der Grient (2011) and De Haan & Krsinich (2014). Chessa, Verburg, & Willenborg (2017) provide an overview of the current index methods used with scanner prices. This work is also related to a large body of literature about CPI methodologies. The most comprehensive study of theory and practice about the CPI can be found in the Consumer Price Index Manual published by the International Labour Office (2004). However, some papers focus specifically on quality changes and clothing, such as Brown & Stockburger (2006), Groshen, Moyer, Aizcorbe, Bradley, & Friedman (2017). The paper is also related to a growing field of literature using online prices to research diverse topics. Lünnemann & Wintr (2011), Cavallo (2013), and Gorodnichenko & Talavera (2017) study the relationship between online and offline markets. Additionally, Harchaoui & Janssen (2018) forecast inflation rates using online price indices.

This paper is organized as follows. Section 2 describes the dataset used, and Section 3 showcases the downward trend typically seen in existing online price indices. Section 4 describes the closest-match method, and Section 5 estimates an online price index for Germany, the Netherlands, Spain, the United Kingdom, and the United States. Section 6 provides the conclusion.

2. The data

This section describes the datasets used to calculate the inflation indices with online prices in Germany, the Netherlands, Spain, the UK and the US from January 2015 to December 2017.

I am interested in calculating an online index and comparing it with the non-seasonally adjusted Consumer Price Index for garments, all urban consumers, calculated by each country's national statistical office.

I use online prices provided by *PriceStats*, a private company that spun off from the Billion Prices Project (BPP) at MIT[1]. The database has been designed to calculate price indices across countries. An advantage over other price databases is that each price is collected along with detailed information about the posted product, such as its name, brand, model characteristics, and out-of-stock indicator. Based on this information, *PriceStats* classifies each product following the United Nation's Classification of Individual Consumption According to Purpose. Most statistical offices use a similar classification structure, so the products included in this paper, which are all classified as garments, overlap with the typical urban consumption basket used in the CPI. A second advantage of this database is that out of stock items are excluded from the database, so the online indices only include items that can be purchased by consumers at each point in time. Online prices include the VAT tax but exclude delivery fees, equalizing the price components consumers pay in the online and offline markets.

There is a growing body of academic research showing that online prices resemble their offline counterparts. For example, Cavallo (2017) collects online and offline prices simultaneously from large multi-channel retailers in 10 countries and documents a high degree of similarity between their price levels. Additionally, price changes were found to occur with similar frequency and to be of similar average sizes in both locations. Aparicio & Bertolotto (2017) present further evidence of the high correlation between online and offline indicators and find that the movements in online price series anticipate movements in the headline CPIs of Australia, Canada, France, Germany, Greece, Ireland, Italy, the Netherlands, the United Kingdom, and the United States.

---

[1] See http://bpp.mit.edu and Cavallo & Rigobon (2016) for additional details on the BPP and the scraping methodology.

Table 1 shows the time-span, the daily average number of items, and the number of sources for each country. *PriceStats* does not disclose the names or details of the sources of its information. However, every source is representative of each country's retailer sector and sells in both online and offline stores.

Table 1 – Data Description

Country	Time Span	Sources	Items per Day
Germany	January 2015 – December 2017	2	9,714
Netherlands	January 2016 – December 2017	2	10,139
Spain	July 2015 – December 2017	1	6,307
UK	July 2016 – December 2017	2	4,791
USA	January 2016 – December 2017	3	5,936

Notes: "Items per day" stands for the average number of items per day and country. The "Sources" is the number of retailers included in any particular country.

Arguments explaining the deflationary tendency of offline clothing price indices have rested on the sector's dynamism and the consistently decreasing nature of garment prices (see De Haan & Hendriks (2013) and Krsinich (2016)). Figure 1 suggests that such notions are true for the database used in this paper. Panel A shows the distribution of the product's lifespans, calculated as the number of days between the first and last observation for each product. More than 60 percent of the products last less than six months, and 90 percent are sold online for less than nine months. This result is consistent with Bascher & Lacroix (1998)'s paper, who report that around 100 percent of the apparel products included in the French CPI are forcedly replaced every year. Panel

B shows the distribution of price change sizes, which indicates that around 75 percent of the changes are price decreases[2].

Figure 1 – Product's Lifespan and Size of Price Changes

Panel A – Distribution of the lifespan of products   *Panel B – Distribution of the size of Price Changes*
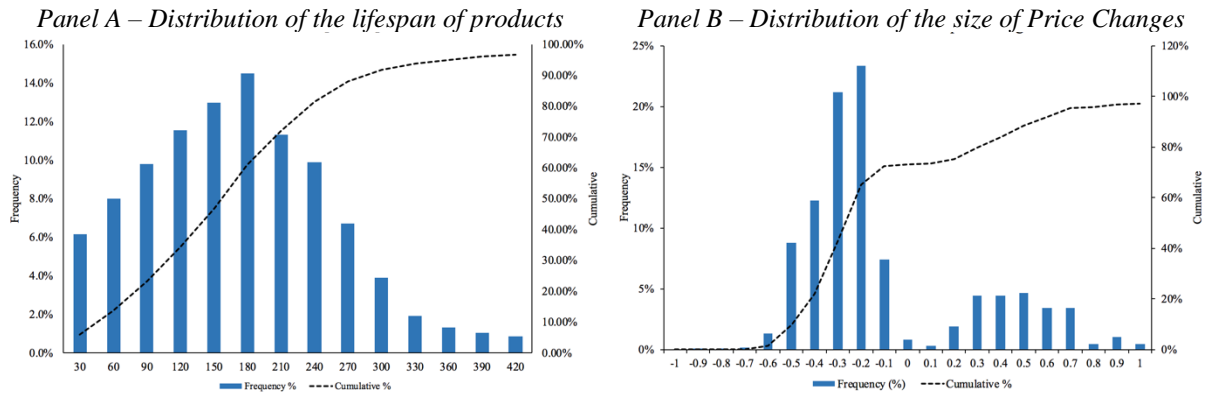
Notes: This figure shows the frequency and cumulative distribution of the product's lifespan and size of price changes. Lifespan is calculated as the number of days between the first and last observation for each product. For any particular product, the size of a price change is calculated as the percentage change of the price between two consecutive observations.

3. The product-turnover problem

To get an intuition for how the CPI collection methodology works, imagine that two products included in the CPI are discontinued from the market in the same month. If the survey agent at a statistical office identifies a comparable replacement for each good, then the next month's CPI should report the price changes comparing the discontinued items and the new products.

Panel A in Figure 2 showcases this situation. Items A0 and A1 are comparable replacements, and the same is true between B0 and B1. Prices are collected for ten months. Assuming that the 2-item

---

[2] One caveat is worth mentioning. This graph does not account for the price increases consumers face when a new product replaces its previous model, which is usually sold at clearance price. As a result, Panel B might underrepresent the number of price increases.

price index follows a Jevons (JV) formula as in the standard CPI methodologies, the ten-month

change in the price index is[3]:

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

Where the $p_{i,t}$ is the price of product i in month t. Intuitively, Equation (1) is the geometric average

of the cumulative price changes of the two qualitatively different items collected by the statistical

office.

Figure 2 – Index Behavior in the Apparel Sector
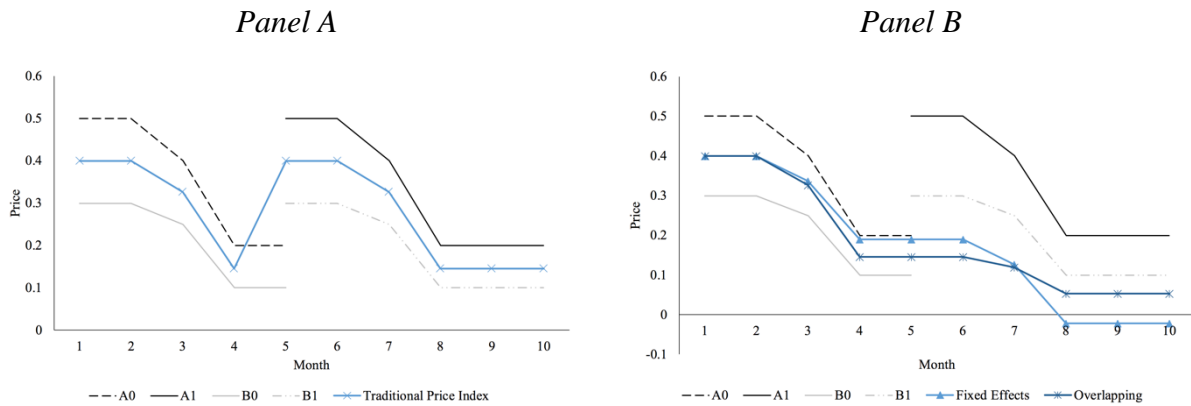
Panel A   *Panel B*

Notes: Panel A shows the typical pricing behavior of four apparel products and a Jevons price index calculated

using these products. Panel B shows the overlapping-quality and fixed-effects price indices calculated based on the

same products.

Now consider the methodologies used to calculate an online price index. Panel B on Figure 2

shows that the online price indices using both the overlapping-quality and the fixed effect (FE)

methodologies lead to an unusual downward trend. The problem is that these methodologies

---

[3] See International Labour Office (2004) for details.

assume that any price difference between the new and obsolete model reflects a quality difference. Appendix A1 shows that this unusual trend is also seen in a real-world data example.

The fixed-effect index disserves special attention because the intuition suggests that it should disentangle the price changes from the quality differences. A mathematical example proves that this is not always the case, so the method cannot be reliably used in sectors such as apparel. Assume that the price $p_{i,t}$ follows a log-linear hedonic model such that:

$$\ln(p_{i,t}) = \alpha + \sum_{t=2}^{T} \delta_t D_{i,t} + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_{i,t} \qquad (2)$$

Where $D_{i,t}$ is the dummy variable with value 1 when product i is present on period t and zero otherwise, and D_i is the dummy variable with value 1 for item i and zero otherwise. The γ_i parameter captures the fixed effect for each item, and δ_t reflects the inflation rate from month one to t. To keep consistency with the graphical example and maintain the results as intuitive as possible, I set $N = 4$ and $T = 10$. The rest of the features of the example (e.g., the time span of each item) also remain unchanged.

Solving Equation 2 by OLS, the cumulative change of the price index in the tenth month is:

$$\Delta P^{FE} = \left(\frac{p_{A1,10}}{p_{A0,1}} \frac{p_{B1,10}}{p_{B0,1}} \right)^{\frac{1}{2}} \left(\frac{p_{A0,5}}{p_{A1,5}} \frac{p_{B0,5}}{p_{B1,5}} \right)^{\frac{1}{2}} \qquad (3)$$

The first term equals Equation 1, while the second term captures the relative prices between the last observation of the old and the first observation of the new models of the goods. If the new

models are more expensive than the old, which is usually the case for the apparel sector, the second term is lower than one. Consequently, the fixed-effect methodology leads to an undesirable downward trend.

Interestingly, if we were able to specify in the fixed-effects model which items are comparable (that is, if we constraint the model such that $\gamma_{A0} = \gamma_{A1}$ and $\gamma_{B0} = \gamma_{B1}$), the price index change on Equation 3 equals Equation 1^4. This result further motivates a possible solution to the downward trend typically seen in the online indices. Online methods should identify comparable items.

4. The Closest match method

The method matches each newly introduced item with its closest alternative good from the existing pool of products. It then compares the first price of each new item against the last price of its closest alternative good and records this as the new item's first price change.

The selection of each product's closest-match is a 2-step process. First, a set of rules filters out most products in the database, identifying a group of similar candidate items. Such rules help reduce the number of computations required in the next step. Second, a formula identifies which of those candidates is the closest match.

The set of rules used in this paper are:

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.

---

[4] See Appendix A2 for further details.

The goal of the filter is to identify items from the previous season, and to ensure that those items are regularly present, avoiding special products that are only present in the market for a very short period. While this paper uses a single set of rules, it is beneficial to set the rules by retailer and sector.

Figure 3 exemplifies which items comply with the set of rules. Product E is introduced into the market on t, so the filter looks for close-alternative goods that were in the database before this date. Item D is not a possible match since it is likely to be from the same season – it was introduced into the market less than 90 days before E's first date. Similarly, item A exited the market more than one year before the introduction of E and it is likely two seasons old, so it is filtered out from the pull of candidates. Possible close-alternative items are B and C^5.

---

[5] Note that item C can still be on the website when the new good enters the index. If C is the closest-match product, then E replaces C. That is, C should be removed from the price index when the new model is introduced. Failing to do so would probably introduce an attenuation bias in the price index.

Figure 3 – Identification of Possible Closest-Match Products
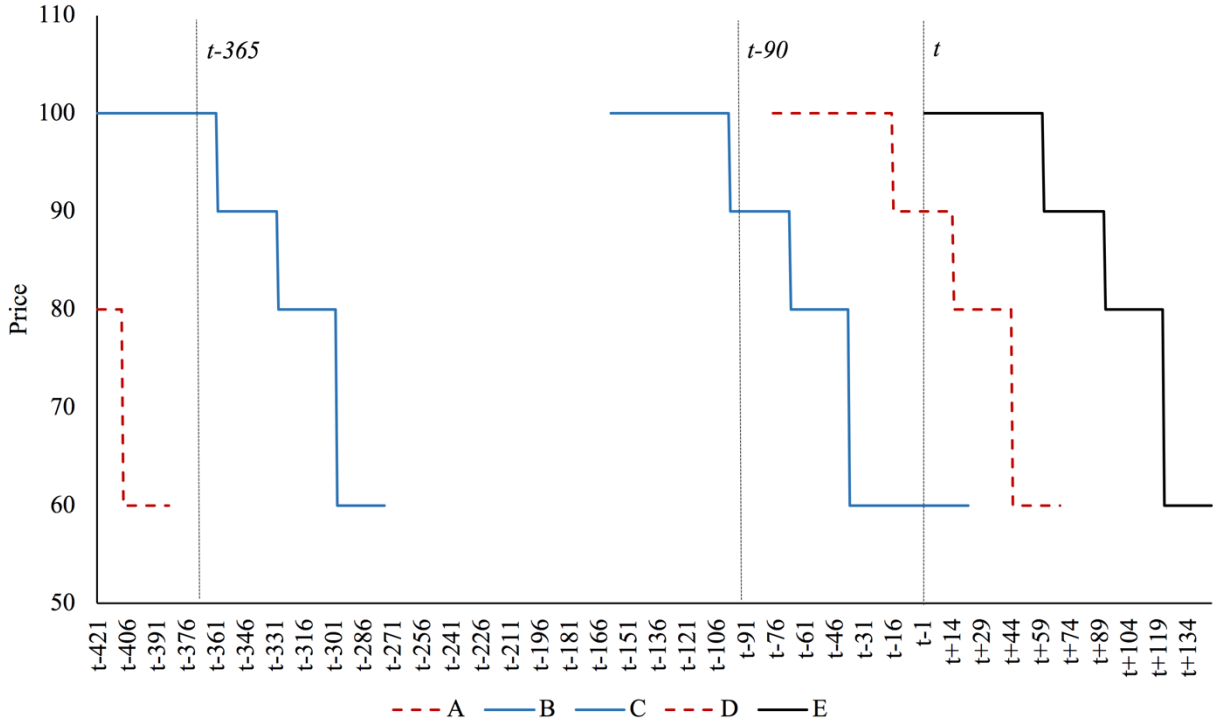
Notes: This figure exemplifies which items can or cannot be considered closest-matches for product E. Product A has not been in the sample for too long and Products D was included in the sample too recently, so none of those are possible matches. In contrast, products B and C are possible matches because those were included in the sample more than ninety days before t, and have been in the sample in the last year.

The second step of the process calculates a score for each feasible close-alternative good. The product with the highest score is considered the closest match. This paper computes the score, $S(q,d)$, using Elasticsearch, a search engine that computes this metric using the following equation:

$$S(q,d) = r(q,d) * \sum_{w=1}^{N} wf(w \; in \; d) \cdot idf(w)^2 \cdot fln(w,d) \qquad (4)$$

11

Where $r(q,d)$ is the relevance factor of the product description of the newly introduced item q and the product description of the feasible close-alternative good d. From now on, d will be referred to as "description" and q as "query". This factor rewards descriptions that contain a high number of query words. The relevance factor multiplies the score by the number of matching words in the description and divides it by the total number of words in the query. The more clauses that match, the higher the degree of overlap between the search request and the descriptions that are returned.

$wf(w\ in\ d)$ represents the word frequency of word w in the product description. This factor assigns a higher relevance to descriptions that repeat a word twice or more. The author recommends setting this value to one since our goal is to find cases where the word is in the description, irrespective to the number of occurrences of the word. Also, retailers do not tend to repeat relevant words.

$idf(w)$ is the inverse description frequency of word w. It is the logarithm of the number of product descriptions in the set of products, divided by the number of descriptions that contain the word. The more often a word appears in the descriptions in the set of products, the *lower* the weight of the word. Common terms like "and" or "the" contribute little to relevance, as they appear in most documents, while uncommon terms like "Nike" or "Adidas" help us zoom in on the most interesting documents.

$fln(w,d)$ is the inverse square root of the number of words in a product description. The shorter the field, the higher its weight. If a word appears in a short description, it is more likely that the content of that description is about the word than if the same term appears in a much bigger body field.

Two additional comments are worth mentioning about the closest match methodology. First, the replacement item can be used more than once. In other words, the same item can be the closest match for two new products. If we only use the replacement item once, we risk using a second-best option as an alternative for one of the new observations.

Second, this methodology reflects a change in paradigm for how old and new varieties of products are linked. While the traditional approach looks for a replacement when an item is discontinued, the closest-match approach searches for a replaced item every time a new product enters the market. Online indices using the traditional approach would present two drawbacks. First, the best match for an item being discontinued may not have arrived in the market yet. In this situation, it is operationally difficult to identify the closest alternative good. Second, new items would not be included in the index from their introduction to the market, but rather when an old product is discontinued. As a consequence, price spells would not be uncensored (Cavallo & Rigobon (2016) states that this is a desirable feature when calculating an online price index).

5. The Online Price Index Versus the CPI

This section compares the traditional CPI to the closest-match online price index. I calculate each country's index as an unweighted geometric average of the price relatives. Details on the methodology can be found in Appendix A3.

As explained in Section 4, this paper calculates a score for each close-alternative good. When this score is lower than a pre-defined threshold, the alternative good is considered significantly different to the new product. In this situation, the new item's price for that period is assumed to

change by the average price change of comparable[6] items. This approach, called class-mean imputation, is currently used in the apparel sector by the Bureau of Labor Statistics (see Brown & Stockburger (2006)).

Every website in the sample displays product information differently. Some retailers list the characteristics of each item, and others describe them in richer text formats. Also, some retailers describe their products in more detail than others. As a result, the score distribution for each retailer is different. Choosing a single threshold for the entire dataset would link goods with significant quality differences in some retailers and would avoid linking products that should be considered replacements in others. Therefore, this paper chooses a unique threshold for each retailer[7,8].

Figure 4 shows the closest-match indices versus the garments CPI. The main conclusion from this figure is that both methodologies yield remarkably similar results. For example, the online Spain index started increasing in September 2015. Three months later, it had already risen by 29.5 percent; similarly, the official release from the national statistical office registered a 30 percent cumulative increase in the same time period. The online price index started falling in January 2016, and the CPI showed a similar change in trend when the estimates for January were published on February 15th. Similar seasonal patterns can be seen in every country.

The Netherlands shows the most considerable discrepancies between the two methodologies. In particular, the online index does not fully capture the price increases after June of 2017. The main reason for this discrepancy is that the number of items in the dataset decreases abruptly (this can

---

[6] Items within the same retailer and category.

[7] The author recommends the following procedure to select the threshold. Select a random set of items and find their closest match. Sort the list of products by the score. 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.

[8] Since the start date of the replaced items is earlier or on $t - 90$, the method requires at least 90 days of data to find closest match products reliably.

be seen in Appendix A4). The closest-match method exploits the large variety of products offered in online stores. Since online databases collect every product sold on a website, it is highly probable that the model will find the previous version or a closely-related version of an item recently introduced into the market. After a few months collecting a small subset of the items sold on the website, the algorithm has fewer alternative items to choose from, and it is harder to find comparable goods. Consequently, only a small set of new products are matched, and the index starts showing the downward bias explained in Section 3.

Figure 4 – Closest-Match Index versus CPI
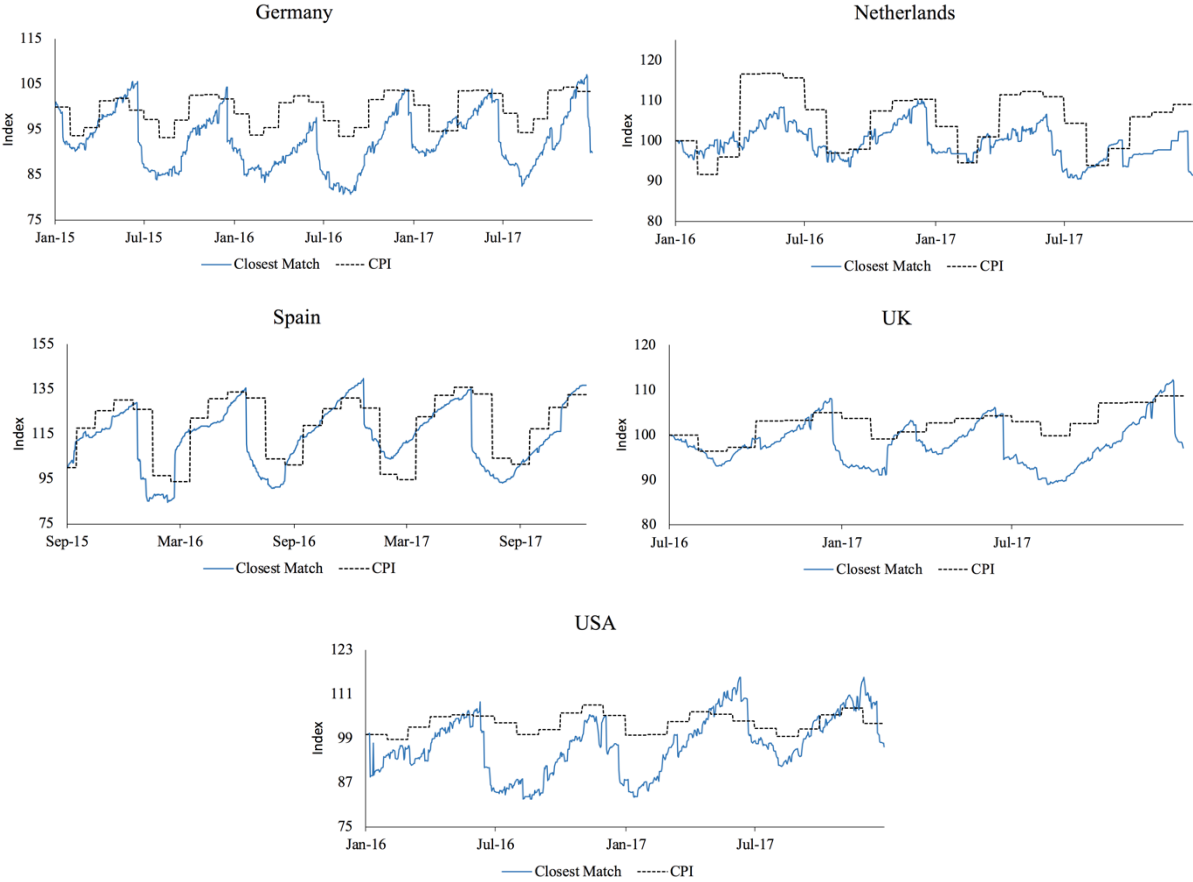
Notes: This figure showcases the closest-match online price index next to the garments Consumer Price Index for each country.

Other small deviations exist between the online price indices and the CPI. The online price indices are sometimes more volatile than their offline counterparts. In line with this view, Gorodnichenko & Talavera (2017) argue that online prices are more flexible than offline prices, and Bertolotto (2016) finds that online prices adjust faster than offline prices to movements in the nominal exchange rate. Additionally, this paper uses a limited number of retailers per country, so it is expected for the CPI estimations to be more robust than the online index to idiosyncratic movements of a particular retailer.

Furthermore, the online price index treats forced replacements differently than the CPI. The set of comparable items identified by a group of specialized agents at the NSO might differ from the set calculated by the closets match method. For those items that are classified as noncomparable, the closest-match method uses a class-mean imputation. In contrast, the CPI typically uses more complex methodologies (such as hedonic regressions).

Taking everything into account, the methodological and sampling differences reinforce the notion that the closest-match price index yields similar results to the traditional CPI.

6. Conclusions

This work introduces a methodology to calculate an online price index that effectively mitigates the effects of product turnover. The method mimics the decision-making process of the specialist that reviews forced replacement items at statistical offices. When a new item is introduced into the price index, the previous version of the item is automatically removed. The method is scalable, so analysts calculating a price index can increase the number of items without significantly increasing the manual effort involved in handling forced replacements.

This paper shows that the closest-match method yields very similar results to the traditional CPI in the apparel category in Germany, the Netherlands, Spain, the United Kingdom, and the United States. For policymakers and anyone interested in inflation dynamics and real gross domestic product estimations, this result contrast with previous online price index results and suggests that the traditional CPI estimates are not biased. Furthermore, the study discusses why current methodologies to calculate the online price index show a downward trend.

This paper suggests areas that would benefit from further research. New research avenues should investigate ways to improve the closest-match method explained in this paper. For example, should we calculate the closest-match score using a different formula? What is the lowest number of product characteristics necessary for the method to accurately decide whether two items are comparable or not? The answers to these questions will depend on the sector where this method is applied.

References

Aparicio, D., & Bertolotto, M. (2017). *Forecasting Inflation with Online Prices*. Retrieved from
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740600

Bascher, J., & Lacroix, T. (1998). *On the use of hedonic methods in the CPI: an application to consumer durables and apparel*. Retrieved from
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+4/$file/1998 4th
Meeting - Bascher & Lacroix - On the use of hedonic methods in the CPI an application to
consumer durables and apparel.pdf

Bertolotto, M. (2016). *The Perils of Using Aggregate Data in Real Exchange Rate Estimations*.
Retrieved from https://ssrn.com/abstract=2882339

Brown, C., & Stockburger, A. (2006). Item replacement and quality change in apparel price
indexes. *Monthly Labor Review*, *129*(12), 35–45.

Cavallo, A. (2013). Online and official price indexes: measuring Argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152–165.

Cavallo, A. (2017). Are online and offline prices similar? Evidence from large multi-channel
retailers. *American Economic Review*, *107*(1), 283–303.
https://doi.org/10.1257/aer.20160542

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for
measurement and research. *The Journal of Economic Perspectives*, *30*(2), 151–178.

Chessa, A. G., Verburg, J., & Willenborg, L. (2017). *A Comparison of Price Index Methods for*

Scanner Data. Retrieved from

https://www.bundesbank.de/Redaktion/EN/Downloads/Bundesbank/Research_Centre/Conf

erences/2017/2017_05_10_ottawa_group_07_1_paper.html?__blob=publicationFile

De Haan, J., & Hendriks, R. (2013). *Online Data, fixed effects and the construction of high-frequency price indexes*. Retrieved from

https://www.researchgate.net/publication/303486094

De Haan, J., & Krsinich, F. (2014). Scanner Data and the Treatment of Quality Change in

Nonrevisable Price Indexes. *Journal of Business and Economic Statistics*, *32*(3), 341–358.

https://doi.org/10.1080/07350015.2014.880059

De Haan, J., & Van Der Grient, H. A. (2011). Eliminating chain drift in price indexes based on

scanner data. *Journal of Econometrics*, *161*(1), 36–46.

https://doi.org/10.1016/j.jeconom.2010.09.004

Goolsbee, A., & Klenow, P. (2018). Internet Rising, Prices Falling: Measuring Inflation in a

World of E-Commerce. *AEA Papers and Proceedings*, *108*, 488–492.

Gorodnichenko, Y., & Talavera, O. (2017). Price Setting in Online Markets: Basic Facts,

International Comparisons, and Cross-Border Integration. *The American Economic Review*,

107(1), 249–282.

Groshen, E. L., Moyer, B. C., Aizcorbe, A. M., Bradley, R., & Friedman, D. M. (2017). How

Government Statistics Adjust for Potential Biases from Quality Change and New Goods in

an Age of Digital Technologies: A View from the Trenches. *Journal of Economic

Perspectives*, *31*(2), 187–210. https://doi.org/10.1257/jep.31.2.187

Harchaoui, T. M., & Janssen, R. V. (2018). How can big data enhance the timeliness of official statistics?: The case of the U.S. consumer price index. *International Journal of Forecasting*, *34*(2), 225–234. https://doi.org/10.1016/j.ijforecast.2017.12.002

International Labour Office. (2004). *Consumer price index manual: theory and practice* (Vol. Geneva). Retrieved from http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf

Krsinich, F. (2016). The FEWS index: Fixed effects with a window splice. *Journal of Official Statistics*, *32*(2), 375–404. https://doi.org/10.1515/JOS-2016-0021

Lünnemann, P., & Wintr, L. (2011). Price Stickiness in the US and Europe Revisited: Evidence from Internet Prices. *Oxford Bulletin of Economics and Statistics*, *73*(5), 593–621.

Appendix

A1. Online price index using the overlapping-quality methodology

Figure 5 shows the online price index using the overlapping-quality methodology versus the garments CPI. This figure demonstrates with real-world data the unusual downward trend explained in Section 3. Although not shown in this paper, the price index using the fixed-effects methodology yields a similar trend.

Figure 5 – Overlapping-Quality Index versus CPI

USA
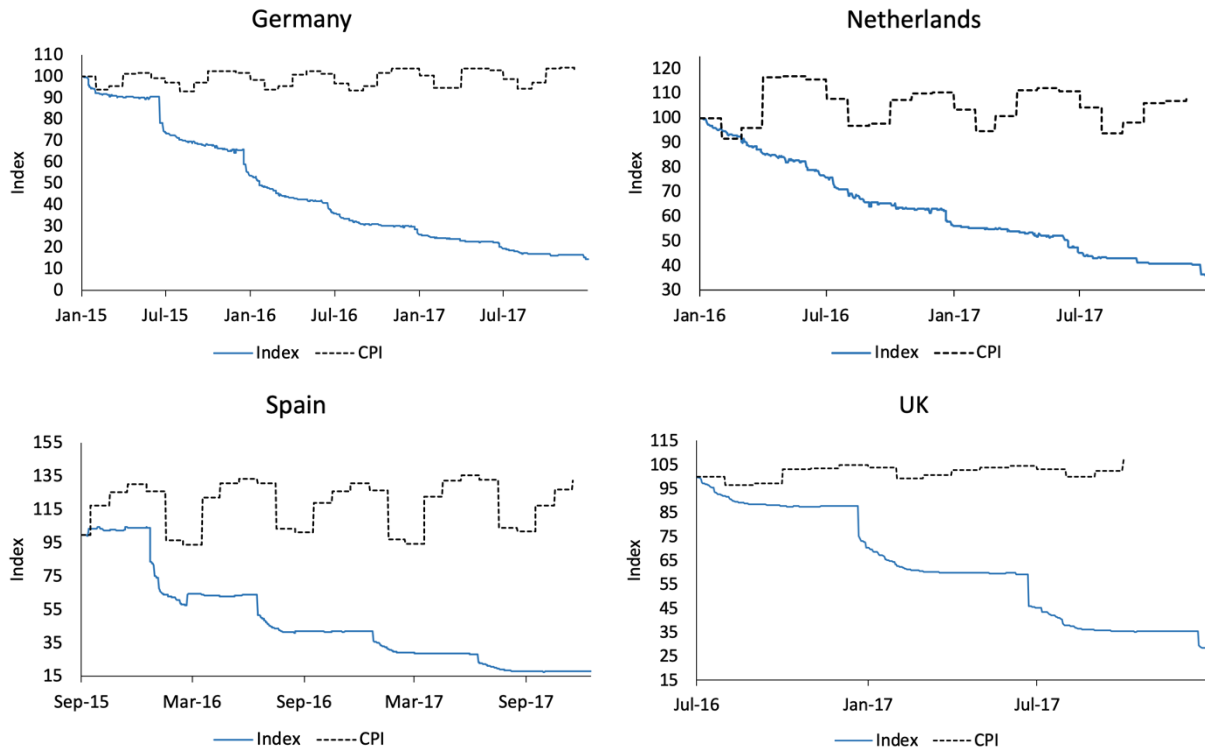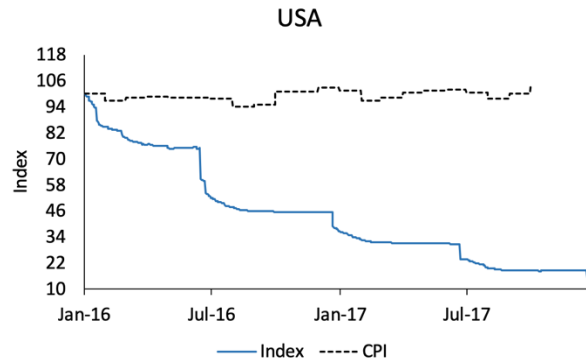
Notes: This figure shows the overlapping-quality online price index next to the garments Consumer Price Index for each country.

A2. Fixed-Effects Example

This section solves for Equation (3), and shows that restricting the fixed-effects coefficients corrects its bias. The matrix representation of the example described in Equation (2) is:

$$
\underbrace{\begin{pmatrix}
\log p_{A0,1} \\
\log p_{A0,2} \\
\log p_{A0,3} \\
\log p_{A0,4} \\
\log p_{A0,5} \\
\log p_{A1,5} \\
\log p_{A1,6} \\
\log p_{A1,7} \\
\log p_{A1,8} \\
\log p_{A1,9} \\
\log p_{A1,10} \\
\log p_{B0,1} \\
\log p_{B0,2} \\
\log p_{B0,3} \\
\log p_{B0,4} \\
\log p_{B0,5} \\
\log p_{B1,5} \\
\log p_{B1,6} \\
\log p_{B1,7} \\
\log p_{B1,8} \\
\log p_{B1,9} \\
\log p_{B1,10}
\end{pmatrix}}_{Y}
=
\underbrace{\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}}_{X}
\cdot
\underbrace{\begin{pmatrix}
\alpha \\
\gamma_{A0} \\
\gamma_{A1} \\
\gamma_{B0} \\
\delta_2 \\
\delta_3 \\
\delta_4 \\
\delta_5 \\
\delta_6 \\
\delta_7 \\
\delta_8 \\
\delta_9 \\
\delta_{10}
\end{pmatrix}}_{b'}
$$

Solving by OLS, Equation (3) is the last row in the vector of coefficients $\widehat{b'} = (X'X)^{-1}X'y$.

Now, I show that constraining the model such that $\gamma_{A0} = \gamma_{A1}$ and $\gamma_{B0} = \gamma_{B1}$, the cumulative change of the price index in the tenth month equals Equation (1). In matrix form, the regression constraints are:

$$
\begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \underbrace{\begin{pmatrix} \alpha \\ \gamma_{A0} \\ \gamma_{A1} \\ \gamma_{B0} \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \\ \delta_8 \\ \delta_9 \\ \delta_{10} \end{pmatrix}}_{b'} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{r}
$$

with the first matrix labeled R.

Solving by constrained OLS, Equation (1) equals the last row in the vector of coefficients:

$$
\hat{b'} = (X'X)^{-1}X'y + (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}\{r - R(X'X)^{-1}X'y\}
$$

A3. Closest-Match Price Index Calculation

This section explains how to calculate the closest-match price index.

1. Load the panel dataset into a statistical package such as R, STATA, Python, or MATLAB.

2. Drop items that appear for only two days or less.

3. Carryforward missing prices for three months.

4. Identify new observations. In other words, tag the non-missing observations that had a missing value on the previous day.

5. For any new item, find its closest alternative good. For this paper, I uploaded the information to Elasticsearch, a search engine that can be customized to filter and calculate the score of each product.

6. Input the price of the closest alternative good one day before the first price of the new item.

7. Calculate the price change per item. Note that step 6 ensures comparing the first price of each new item against the last price of its closest alternative good and records this as the new item's first price change.

8. Remove price relatives higher than 10 and lower than 0.1. This is a recommendation from the Consumer Price Index Manual published by the International Labour Office (2004), chapter 9.

9. Calculate the geometric mean of price changes.

10. For those new items with no close alternative good, assume that their initial price change is the average of the price changes of similar items on that day. This is referred to as Class-mean imputation in Brown & Stockburger (2006).

11. Calculate a price index based on the average price changes calculated in the previous steps.
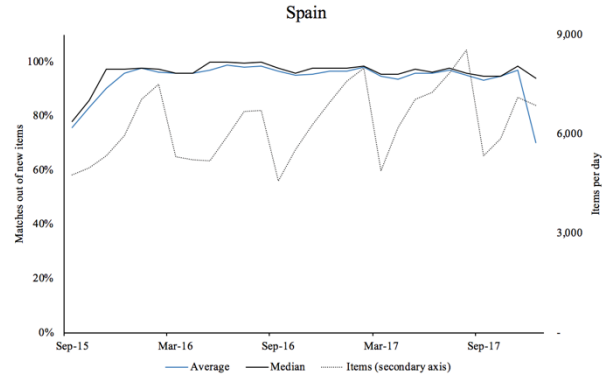
A4. Robustness of The Online Price Indices

Figure 6 shows the average and median number of goods successfully matched to new items relative to the total number of new products, per month. The secondary axis shows the average number of items per month.

Two observations are worth mentioning about the robustness of the online price indices calculated in this paper. First, the number of items in the garments online index is highly volatile. This is a typical feature of the clothing data, where online stores add new product models to their websites

but do not discontinue their previous models on that same day. Some items are discontinued either before or after the end of the season. Second, such volatility does not impact the number of matched items per month as a percentage of the new items, which is stable. Taking into account the results on Figure 4, it is therefore reasonable to suggest that the volatility in the number of items does not show a severe impact on the online indices either.

Figure 6 – Products with Close-Alternative Goods

Spain
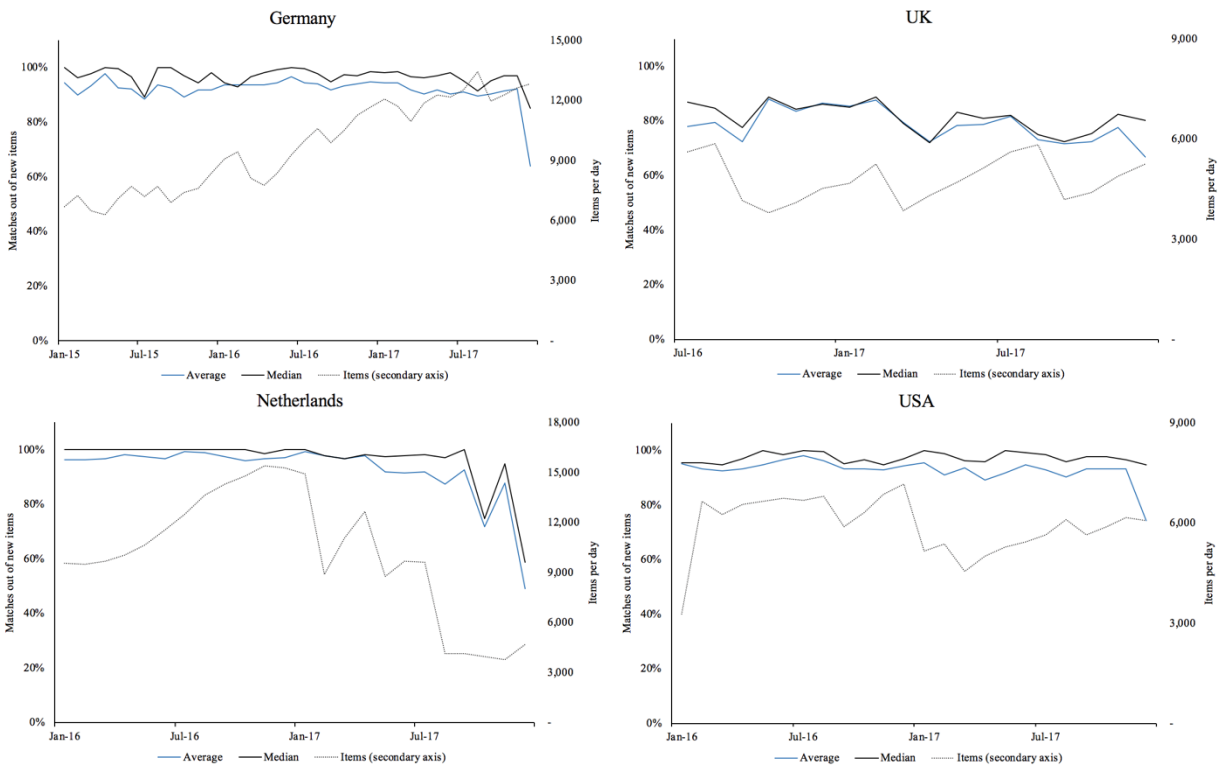
Notes: This figure shows the average and median of the number of goods successfully matched to each new item relative to the total number of new products per month, for each country. The secondary axis shows the average number of items per month.

Acknowledgment