Price Ind[1]ex for Clothes and Shoes in Finland: Solving the Life-Cycle Bias in Scanner Data

Suoperä, Antti, Statistics Finland,

Luomaranta, Henri, Statistics Finland,

Nieminen, Kristiina, Statistics Finland,

Markkanen, Hannele, Statistics Finland

Abstract

In a setting with full enumeration of price items on clothes and shoes, pairwise comparisons or bilateral price links fail to capture the life-cycle effects due to strong attrition, and a new methodology is needed for the quality adjustment. The new scanner datasets allow to identify all the products via unique identifiers, GTIN-codes or VNR-codes and one may follow the products for long periods of time (several years), supporting the construction strategies based on bilateral price links. However, the life cycles of clothes and shoes are too short for these traditional strategies. In this study, we suggest a new method for clothes and shoes, based on a very simple hedonic method. Our method is robust to modeling errors and corrects for the life cycle bias that originates from ignoring the overlap periods. This is demonstrated by using a *complete* scanner data set, which contains monthly observations of prices and quantities from January 2020 to April 2021 and encompasses different types of outlets and regions.

Keywords
Life Cycle of Commodity Bias, Robust Hedonic Method, Conditional Averages, Unit Value, Basic and Excellent Index Number Formulas, Consumer Price Index.

1

EU-grantaggreementno 885851, 2019-FI-B4469-PRICE –New methodologies for HICP

1 Introduction

The data generating process of prices for clothes and shoes is different from daily products. First, in the case of daily consumer goods one may track commodities that are of comparable quality (i.e. GTIN-commodities) for longer periods, i.e. several years, while clothes and shoes can be hardly tracked for six months. This paper shows that the attrition of commodities in clothes and shoes is very strong. Second, the price development of clothes and shoes is different when compared to 'a normal commodity' – in most cases, prices tend to decline and have the lowest price at the end of the life cycle.

These two properties render the usual bilateral index construction strategies less useful. We show that bilateral strategies lead to indices that converge rapidly towards zero when attrition is high. This would suggest that prices of clothes and shoes decline over time. However, this is not a true representation of reality, but rather, the index declines because of misspecified construction strategy. The problem arises at a time period when an old vintage disappears and a new one comes to the market – at so called overlap period. The bilateral strategies are not able to capture this information. The 'HICP recommendations on bridged overlap' (2021) suggest some robust imputation methods to overcome the problem. Instead, we propose two different methods based on a price model, 1) quality adjustment and hedonic index (i.e. Oaxaca decomposition), and 2) a robust compilation strategy based on observed weighted average prices (unit values) that corresponding to the conditional averages – based on age of the product (replicating the basic algebraic property of the OLS).

The structure of the study is as follows. The following chapter 2 shows how serious a problem the one-sided-null-values (i.e. new and disappearing commodities) is for clothes and shoes. Chapter 3 discusses the shortfalls of indices constructed by using bilateral price-links. We analyze an index for commodities that are comparable in quality and compare some basic and excellent index number formulas. In chapter 4 we present the improved methodology - how the index should be constructed using hedonic methods? In this chapter we also present a robust method that is an ideal solution for the overlap problem. In chapter 5 we conclude.

2 Serious Attrition of Clothes and Shoes

The life cycle of range of clothes and shoes is short compared to daily products. This is evident in scanner data obtained from a Finnish retail trade chain. These (big) data enable us to make very detailed partitions of commodities, based on regions according to NUTS2 classification, outlet type, regional outlet, most detailed commodity category and GTIN-identifier. In the following, we track GTIN commodities that are of comparable quality.

As the backdrop, it is useful to note that Finland has four seasons, which in principle, require different types of clothing. However, the autumn and spring are quite similar in weather type, so spring and autumn clothes can be quite similar. This is why we have two main alternatives for the base strategy. We can either consider the entire previous year, or alternatively, only the preceding half of the year as the base. To better capture the effects of the serious attrition of commodities, we have decided to select the preceding half-year period as base and normalize the base to an average month, i.e. we compute the average of the months of the preceding six months. Figures 1 and 2 show how strongly the values of consumption (RSB) and the number of GTIN commodities (RSN) decline during the period of six months. During the relatively short time span, about 60 percent of consumption have disappeared in all NUTS2 regions, which is too much in order to remain representative of consumption patterns.

In Figures 1 and 2 the base (the first six months of the life cycle) includes all price-links, including disappearing and new commodities. The basic question we need to consider, is the rate of decline in consumption from the base, up to the end of the year 2020 (i.e. 2020/7 – 2020/12). The figures 1 and 2 show, that the decline in consumption happens quite fast. Thus, the life cycle for GTIN identified clothes and shoes to be very short in Finland due to four seasons and this leads to very serious attrition of GTIN commodities. This evidence points to the fact that any method based on bilateral price-links is doubtful. The next chapter completes the story by assessing the index construction strategies based on bilateral price-links.

Figure 1: Decline of consumption values (RSB) and number of commodities (RSN) in NUTS2 region West Finland in year 2020.

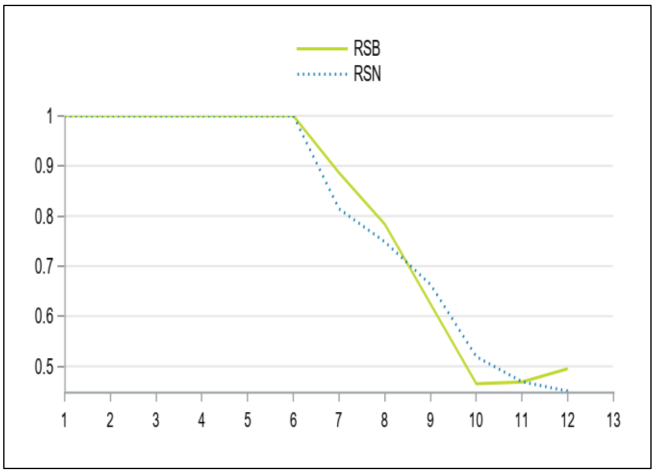
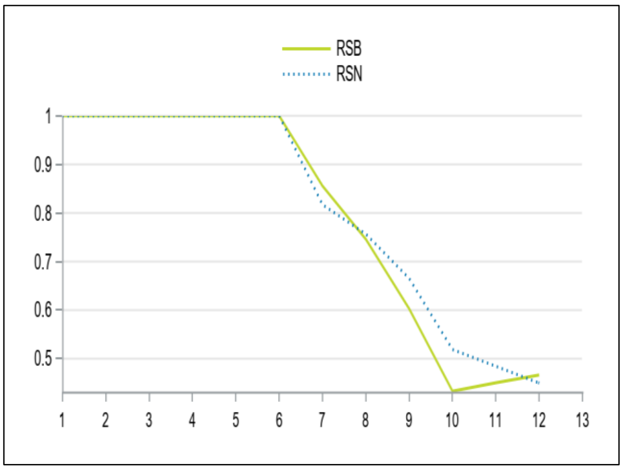


Figure 2: Decline of consumption values (RSB) and number of commodities (RSN) in NUTS2 region Helsinki-Uusimaa in year 2020.



The problem of disappearing commodities is similar for all NUTS2 regions for clothes and shoes.

3 Index Based on Bilateral Price-Links

We can avoid the drift problem by applying a strategy which allows shifting the base as appropriate. We define the preceding half year period as base and change it the first time in July 2020. In practice, this means that our first base includes the first six months from 2020 and the index is calculated by this base for the whole year. The first time we change the base is at January 2021 and the second half of the year 2020 becomes the base. Again, the base is normalized as an average month during the base period.

In the following exercise, we follow GTIN commodities that are comparable in quality. We show that the index based on comparable EAN commodities shows strongly declining prices when attrition is high. This is true for all bilateral methods and strategies – and also for multilateral methods. In practice, this means that the index series converges quickly towards zero.

We use two sets of index numbers (Vartia & Suoperä, 2018):

1. The basic index numbers: L = Laspeyres, LL = Log-Laspeyres, Lh = Harmonic-Laspeyres, Pl = Palgrave, LP = Log-Paasche and P = Paasche.
2. Excellent index numbers: S = Stuvel, T = Törnqvist, MV = Montgomery-Vartia, SV = Sato-Vartia, W = Walsh and F = Fisher

We calculate index numbers separately for clothes and shoes in five NUTS2 regions. The results, presented in Figures 3-6, are troublesome. We notice that the results for all five NUTS2 regions are very similar for clothes and shoes. The following conclusions emerge:

1. The basic index numbers (Figures 3 and 5) are *contingently* biased (see Vartia, Y. & Suoperä, A. 2018), but the excellent ones go ‘hand-in-hand’ (Figures 4 and 6).

2. The index series for GTIN commodities decline very rapidly and converges towards zero (never reaching it), say after few years. This result implies that prices will also converge to zero but is that true and why is this happening? Quite evidently there is something wrong with construction strategy of index based on bilateral price links.
3. The index series for clothes and shoes presents that prices have declined from January 2020 to February 2021 (14 months) by almost 40 %. But, have they really declined?
4. Indices constructed for GTIN commodities by bilateral price-links neglects the problem of 'one-side-null-values' (i.e. disappearing and new commodities) or more specifically the problem of overlap.
5. We would argue that prices of clothes and shoes should not decline but increase instead, at least in the short-run.
6. The problem arises: 'How to solve the index problem of declining prices based on bilateral price-links (i.e. commodities being comparable in quality)?'

Figure 3: Declining prices by the basic index numbers. Index series are 2020=1 and NUTS2 region is 'West Finland' for clothes.

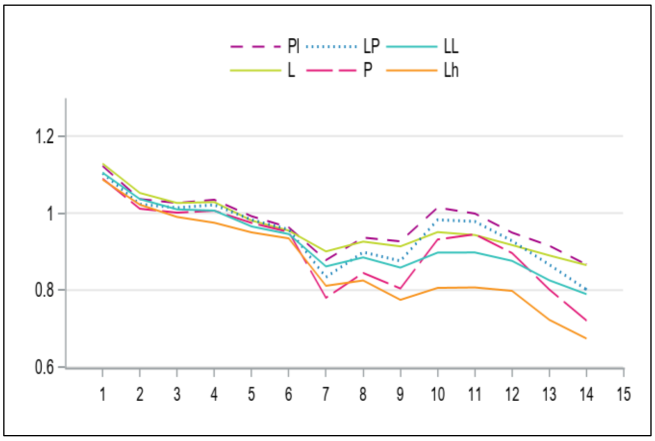


Figure 4: Declining prices by the excellent index numbers. Index series are 2020=1 and NUTS2 region is 'West Finland' for clothes.

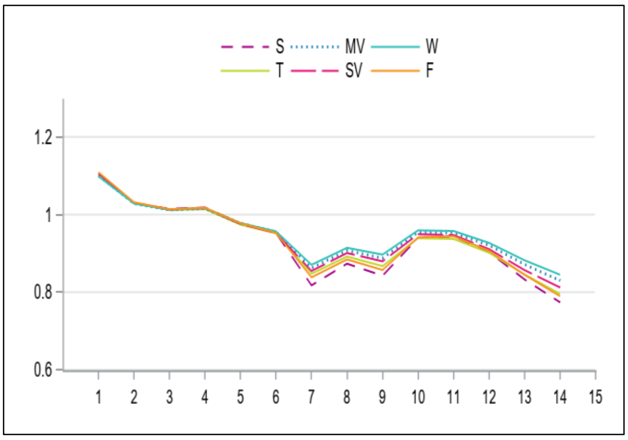


Figure 5: Declining prices by the basic index numbers. Index series are 2020=1 and NUTS2 region is 'West Finland' for shoes.

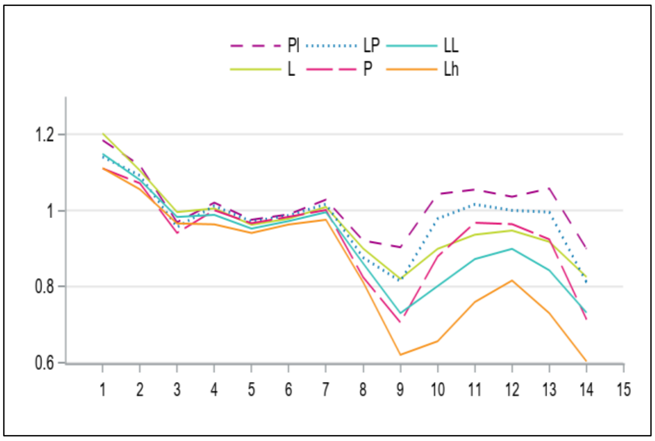
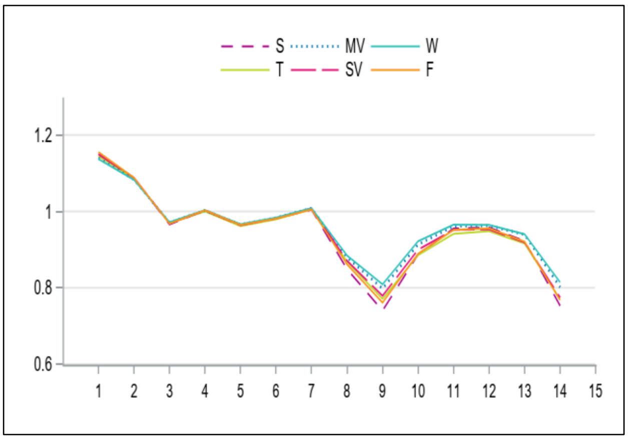


Figure 6: Declining prices by the excellent index numbers. Index series are 2020=1 and NUTS2 region is 'West Finland' for shoes.



The upward bubble at 2020/10 and 2020/11 tells that consumers are preparing for wintertime by buying new, more expensive winter collection which affects prices temporarily.

4 The Problem of Overlap

Let us consider a problem of overlap with a simple example, say ‘T-shirts’. There are old (yellow) and new (blue) ranges of T-shirts. The life cycle of the first range (or perhaps assortment) of T-shirts is 9 months and after that a new set of products comes to the markets and replaces the old ones within a month or two – this is the period of overlap. Both ranges include hundreds of GTIN-commodities but in Figures 7 and 8, they are aggregated into six groups (dotted lines) which are comparable in quality. The solid line represents index series for all commodities. It is constructed by base strategy and by Törnqvist formula (Figure 7). We do the same for their actual average prices (Figure 8). Notice the difference in scaling of y-axis.

Figure 7: Index series for old (yellow) and new (blue) ranges of T-shirts. The solid lines are aggregate Törn-qvist for old and new ranges of T-shirts.

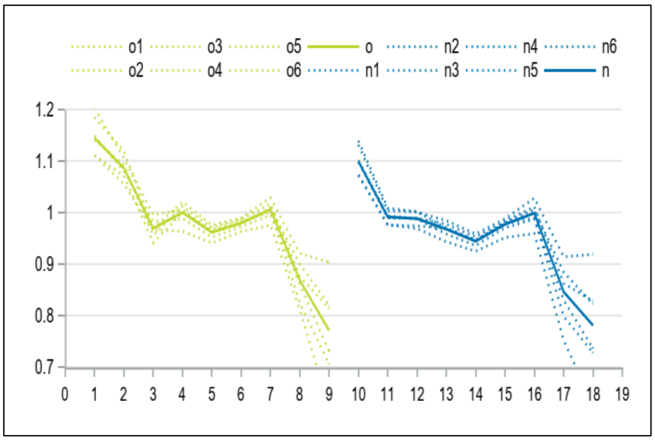


Figure 8: The same by commodity prices. The solid lines represent price profiles for average prices.

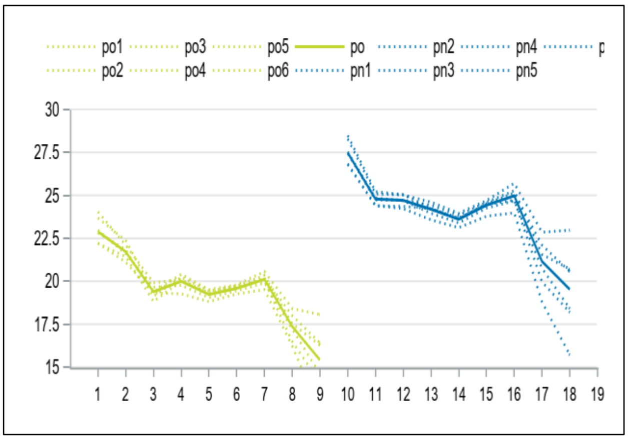


Figure 7 above shows Törnqvist price indices for six groups and for the aggregate category, and Figure 8 above presents the corresponding average price profiles for the same groups respectively. The yellow lines represent the old range and the blue represents the new range. The overlap period is from 9 to 10, when the new range replaces the old one. We learn that:

1. For both sets of ranges index series for commodities and their aggregates declines.
2. Figure 7 tells that prices mostly decline but Figure 8 shows that prices increase on average (the blue lines are at higher level compared to yellow lines). Thus, the figures are an example of a very well-known heterogeneity bias (Hsiao, 1986, p. 5-8) that in our case leads to serious downward bias.
3. Some statistician believes that the price differences from the period 9 to 10 is due to quality change between these two ranges of T-shirts. This is unlikely, given the short time span.

The problem that naturally arises is how to link the yellow and blue lines (index series of two set of T-shirts) to get the index series for the entire time period 1 to 18? HICP recommendation on bridge overlap (Eurostat, 2021) consists of several implementation methods but unfortunately these methods lead to up-, or downward bias (Eurostat, Table pp. 6 or Annex 2, 2021). Next, we propose a new method that is based on conditional averages and quality adjusting. The method is based on estimates that are statistically unbiased.

4.1 The Hedonic Solution for the Problem of Overlap

In deterministic approach of index numbers, a price index depends on prices and quantities from base and observation periods, for example from periods 0 and *t*. For hedonic index this is not enough because prices depend on also k-vector of quality characteristics or variables, say from $\boldsymbol{x}'_{it}$ and in this paper also on a life cycle of a range of commodity, say z_{it}. In practice, this variable measures the age of range of commodities, that is, the number of months the commodity has been on the markets. We think that this is the most important variable to explain the systematic variation of prices. We also think that the age of ranges is an important variable for the solution of the overlap problem. As a clarifying example, we define a simple semilogarithmic price model for two ranges of T-shirts. We specify the price model to be linear in respect to parameters and estimate it by the OLS method. Following the analogy presented in Suoperä & Auno (2021) we aggregate the observation for both ranges and get

(1) $$\log(\overline{p_t}) = \hat{\alpha}_t + \bar{z}_t\hat{\gamma}_t + \bar{\boldsymbol{x}}'_t\hat{\boldsymbol{\beta}}_t$$

where $\log(\overline{p_t})$ is logarithm of unit value (i.e. weighted average price), a vector $\bar{\boldsymbol{x}}'_t$ presents weighted averages of exogenous independent variables (i.e. quality characteristics). Prices depend on the k quality variables, but also from the average 'age of range or design', $\bar{z}_t$. The estimate $\hat{\gamma}_t$ tells us how much prices decline when the design/fashion of commodity increases by one period. Notice that the average model for the most detailed commodity category group does not include the error term because it sums up to zero by definition of the OLS (Suoperä & Auno, 2021).

Some basic properties of equation (1) are:

1. For two sets of T-shirts *separately*, the quality of commodities do not change, that is $\bar{\boldsymbol{x}}'_t = \bar{\boldsymbol{x}}'_\tau$ (ninth to tenth period is the overlap period), but between the range of T-shirts quality may change.
2. For the age of design of the two ranges of T-shirts always holds, that $\bar{z}_t \neq \bar{z}_\tau$, for all $t \neq \tau$ meaning that the age of design is not equal in any two separate time periods. To put it simply, in different stages of life cycle, the commodities will have different ages
3. We say that prices will not decline in the Figure 7 because of quality change, but because of their ages (any given two sets of commodities analyzed *separately* it holds that $\bar{\boldsymbol{x}}'_t = \bar{\boldsymbol{x}}'_\tau$, but for age $\bar{z}_t \neq \bar{z}_\tau$, for all $t \neq \tau$). This can be presented by so-called Oaxaca decomposition as follows:

   $$\log(\overline{p_t}/\overline{p_\tau}) = (\hat{\alpha}_t + \bar{z}_t\hat{\gamma}_t + \bar{\boldsymbol{x}}'_t\hat{\boldsymbol{\beta}}_t) - (\hat{\alpha}_\tau + \bar{z}_\tau\hat{\gamma}_\tau + \bar{\boldsymbol{x}}'_\tau\hat{\boldsymbol{\beta}}_\tau) \leftrightarrow$$

   (2) $$\log(\overline{p_t}/\overline{p_\tau}) = \left\{(\bar{z}_t - \bar{z}_\tau)\hat{\gamma}_t + (\bar{\boldsymbol{x}}'_t - \bar{\boldsymbol{x}}'_\tau)\hat{\boldsymbol{\beta}}_t\right\} + \left\{\hat{\alpha}_t - \hat{\alpha}_\tau + \bar{z}_\tau(\hat{\gamma}_t - \hat{\gamma}_\tau) + \bar{\boldsymbol{x}}'_\tau(\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_\tau)\right\},$$ where

   first term on right side in (2) is the *quality correction of 'design/fashion of age' and quality variables* and the second term in (2) is the *quality adjusted price change evaluated at standard point of quality*, that is at $\bar{z}_\tau$ and $\bar{\boldsymbol{x}}'_\tau$ (see analogy with Suoperä & Auno, 2021).

How to interpret the above equations by graphs? Hedonic index says that the prices of first and second set of T-shirts (i.e. yellow and blue separately) declines because of increased age of the vintage.

Figure 9: Index series for old (yellow) and new (blue) ranges of T-shirts. The red one is quality adjusted Törnqvist.

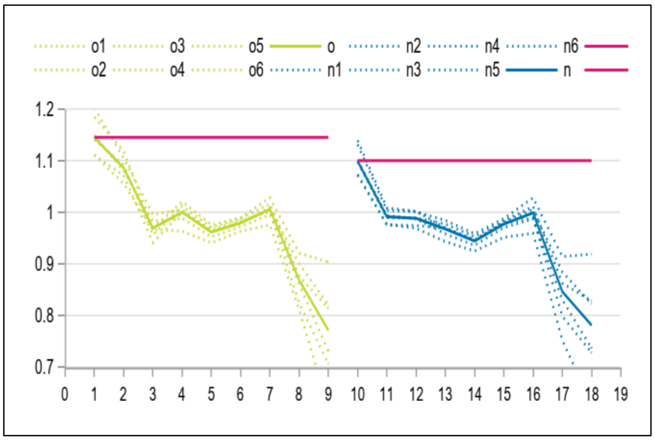


Figure 10: The same by commodity prices. The solid yellow and blue lines represent price profiles for true average prices and red profile quality adjusted prices.

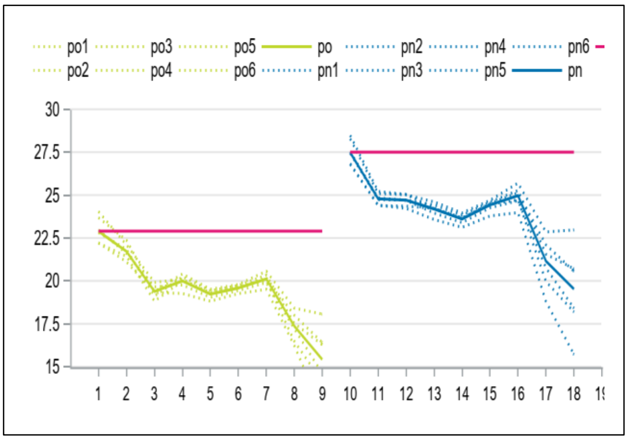


Figure 7 the quality correction lies between two red lines: yellow and red or blue and red lines. This happens because:

1. Yellow case: $\bar{\boldsymbol{x}}'_t = \bar{\boldsymbol{x}}'_\tau$, for all $t \neq \tau$, and t, τ =1,…,9, but $\bar{z}_t \neq \bar{z}_\tau$, for all $t \neq \tau$. The difference between yellow and red lines is exactly the *quality correction of the age of design*, that is $(\bar{z}_t - \bar{z}_\tau)\hat{\gamma}_t$.
2. Blue case: $\bar{\boldsymbol{x}}'_t = \bar{\boldsymbol{x}}'_\tau$, for all $t \neq \tau$, and t, τ =10,…,18, but $\bar{z}_t \neq \bar{z}_\tau$, for all $t \neq \tau$. The difference between blue and red lines is exactly the *quality correction of the age of design*, that is $(\bar{z}_t - \bar{z}_\tau)\hat{\gamma}_t$.

The overlap happens between ninth and tenth period. One way to calculate price change between the last (average) actual price of yellow line (about 15.4 euro) and the first (average) actual price of the blue line (about 27.5 euro), that is 27.5/15.4 ≈ 1.78. We think this method is seriously biased, because:

1. The difference of the age of design/fashion of the ranges between the last price of yellow range and the first price of the blue range is $(9 - 1)$.
2. We think that quality of yellow and blue ranges does not change practically in any way during the short time-span i.e. $\bar{\boldsymbol{x}}'_9 \approx \bar{\boldsymbol{x}}'_{10}$.
3. We suggest removing the quality change of the last price of the yellow range (i.e. $(9 - 1)\hat{\gamma}_t \approx 7.5\text{ euro, or log }(7.5)$) that is the difference between yellow and red lines (=quality adjusted price of the yellow range) and then computing the quality adjusted price change between overlap periods (periods 9 and 10) from red lines, that is: 27.5/22.9 ≈ 1.2 or log(27.5/22.9) ≈ 0.18 log-%.

In Finland, markets for clothes and shoes for the most detailed commodity categories in NUTS2-regions behave as in our example. There are perhaps hundreds of ranges of clothes and shoes that appear to the markets almost at the same time – approximately when their season starts. Their qualities, as measured by the $\bar{\boldsymbol{x}}'_t$ quality variables, are approximately equal over a short time span, but their ages deviate between the base and observation periods. Traditionally the observation period includes the most recent fashion of clothes and shoes meaning that we must adjust the base period prices upward by $(\bar{z}_t - \bar{z}_0)\hat{\gamma}_t$ to obtain the quality adjusted prices for base period. In the following chapter we show how this can be done with scanner data.

4.2 The Robust Solution for the Problem of Overlap

The hedonic method is a model-based approach for index numbers. In the previous chapter we show that the overlap problem can be solved by estimating the quality correction term $(\bar{z}_t - \bar{z}_0)\hat{\gamma}_t$, and by deriving the quality adjusted prices at a standard point of quality, say at $\bar{z}_\tau$ and $\bar{\boldsymbol{x}}'_\tau$. The hedonic method is quite complicated, and we therefore suggest another, more robust method that also solves the overlap problem. Commodities in our data are grouped by very detailed commodity groups (about 1000 groups) and we show an analysis of a hypothetical commodity group as an example. For all commodities in that most detailed commodity group we know the quantity consumed, prices, and the age of range of that commodity. We have all that is needed, so we collect these data in Table 1. Age of range is in columns while prices and quantities for two periods are in rows.

Table 1: Example of robust method in observation level.

Age of range	0	1	2	3	4	…	16
Base period 0, prices	$p_{0,i0}$	$p_{1,i0}$	$p_{2,i0}$	$p_{3,i0}$	$p_{4,i0}$	…	$p_{16,i0}$
Base period 0, quantities	$q_{0,i0}$	$q_{1,i0}$	$q_{2,i0}$	$q_{3,i0}$	$q_{4,i0}$	…	q
Observation period *t*, prices	$p_{0,it}$	$p_{1,it}$	$p_{2,it}$	$p_{3,it}$	$p_{4,it}$	…	$p_{16,it}$
Observation period *t*, quantities	$q_{0,it}$	$q_{1,it}$	$q_{2,it}$	$q_{3,it}$	$q_{4,it}$	…	$q_{16,it}$

We have knowledge of quantities consumed for all prices in base and observation periods and the next step is to aggregate each age group prices into unit prices (i.e. weighted arithmetic averages), as shown in Table 2

Table 2: Example of robust method in aggregate level (i.e. for some commodity group).

Age of range	0	1	2	3	4	…	16
Base period 0, average prices	$\bar{p}_{0,0}$	$\bar{p}_{1,0}$	$\bar{p}_{2,0}$	$\bar{p}_{3,0}$	$\bar{p}_{4,0}$	…	$\bar{p}_{16,0}$
Base period 0, sum of quantities	$q_{0,0}$	$q_{1,0}$	$q_{2,0}$	$q_{3,0}$	$q_{4,0}$	…	$p_{16,0}$
Observation period *t*, average prices	$\bar{p}_{0,t}$	$\bar{p}_{1,t}$	$\bar{p}_{2,t}$	$\bar{p}_{3,t}$	$\bar{p}_{4,t}$	…	$\bar{p}_{16,t}$
Observation period *t*, sum of quantities	$q_{0,t}$	$q_{1,t}$	$q_{2,t}$	$q_{3,t}$	$q_{4,t}$	…	$q_{16,t}$

By information of Table 2 we may calculate price index for commodity group, say $P_f^{t/0}$, where *f* is index number formula, that is

(3) $$P_f^{t/0} = exp\{\sum_{j=0}^{16} w_{j,t} * log(\bar{p}_{j,t}/\bar{p}_{j,0})\},$$

where weights for index formula *f* are presented in Table 3 (see Suoperä, Nieminen, Montonen & Markkanen, 2020).

Table 3: Weights for some excellent index number formulas in logarithmic form (i.e. equation (3)).

Excellent formula (see Suoperä, Nieminen, Montonen & Markkanen (2020): Table 3, p.6):	
Törnqvist, f = T	$w_{j,f} = \bar{w}_{j,T} = 0.5 \cdot (w_j^0 + w_j^t)$
Sato-Vartia, f = SV	$w_{j,f} = \bar{w}_{j,SV} = \frac{L(w_j^t, w_j^0)}{\sum L(w_j^t, w_j^0)}$
Montgomery-Vartia, f = MV	$w_{j,f} = \bar{w}_{j,MV} = L(p^t q^t, p^0 q^0)$
Fisher, f = F	$w_{j,f} = \bar{w}_{j,F} = 0.5 \cdot (L(p^t q^0, p^0 q^0) + L(p^t q^t, p^0 q^t))$
Walsh, f = W	$w_{j,f} = \bar{w}_{j,W} = (w_j^0 \cdot w_j^t)^{1/2}$
Stuvel, f = S, Pursiainen, 2006, p. 88	$w_{j,f} = \bar{w}_{j,S} = L(p^t \bar{q}, p^0 \bar{q})$

Our method removes the price effects of ages of ranges between base and observation period. This is probably the *most ideal index* for clothes and shoes, and similar commodities where the quality characteristics approximately satisfy $\bar{\boldsymbol{x}}'_0 \approx \bar{\boldsymbol{x}}'_t$ meaning that quality correction is insignificant at least for short time-spans. Also, the basic algebraic properties of the OLS method says that the quality adjusted price index based on the equation (2) and its Oaxaca decomposition leads to the $P_f^{t/0}$ if $\bar{\boldsymbol{x}}'_0 \approx \bar{\boldsymbol{x}}'_t$ holds. This method is extremely simple and efficient for commodities having short life cycle. To define the life cycle for commodities reliably, we need scanner data at least from current and previous year, but two previous years is surely desirable.

4.3 Robust vs Bilateral Analysis: Preliminary Results

Our empirical results should be considered preliminary because we have data only from January 2020 to April 2021 (16 months) and the reliable definition of life cycle for all commodities needs more time periods – perhaps two years. We compare bilateral analysis based on commodities of similar quality to our extremely simple ideal robust method. We construct indices by the base strategy, where the base for both methods is the second half of the year 2020 (i.e. from July to December). Indices are constructed for time periods from July 2020 to April 2021 and normalized as P(July 2020)=1. Differences between Figures 3 to 6 and 11 to 14 is caused by different definition of the base. In Figures 3 to 6 the base is the first half of the year 2020 for all months of the year 2020 but in Figures 11 to 14 the base is the end half of the year 2020 for all analyzed months. In the other words, the construction strategy of index series presented in the sets of Figures (i.e. 3 to 6 and 11 to 14) differ from each other's and causes seemingly differences, but anything more.

Figure 11: Index series for bilateral (dashed lines) and our robust method (solid lines) in 'West-Finland' for clothes. $P_f^{t/0}$ is Törnqvist.

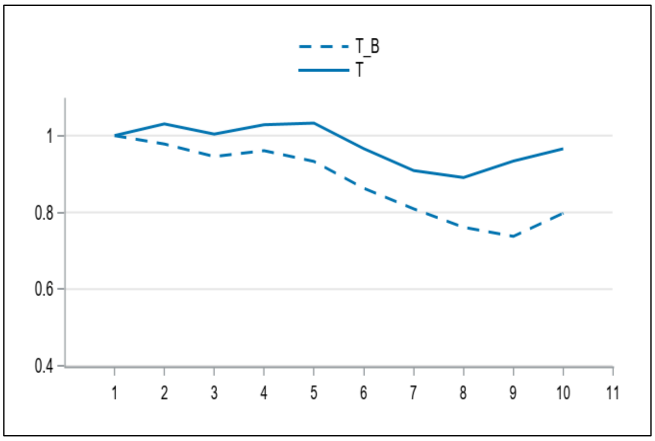


Figure 12: Index series for bilateral (dashed lines) and our robust method (solid lines) in 'West Finland' for shoes. $P_f^{t/0}$ is Törnqvist.

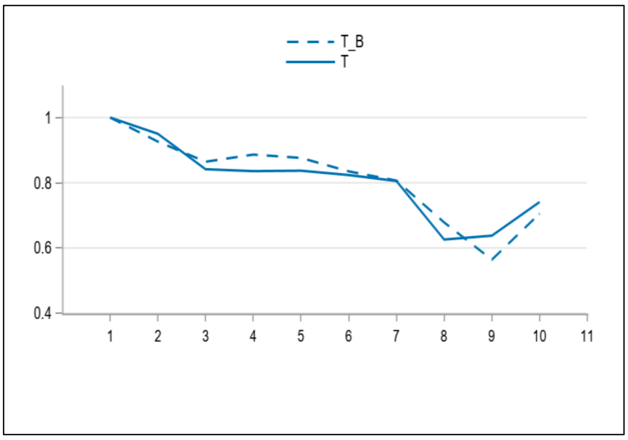


Figure 13: Index series for bilateral (dashed lines) and our robust method (solid lines) in 'Uusimaa' region for clothes. $P_f^{t/0}$ is Törnqvist.

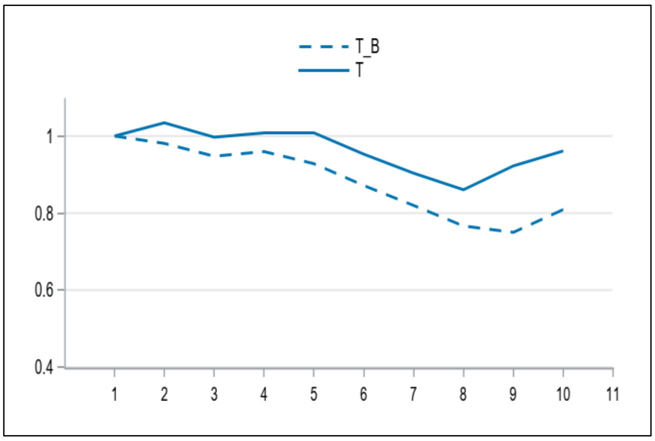
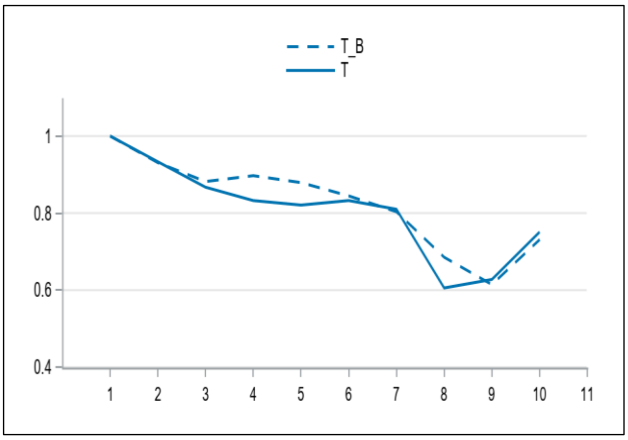


Figure 14: Index series for bilateral (dashed lines) and our robust method (solid lines) in 'Uusimaa' region for shoes. $P_f^{t/0}$ is Törnqvist.



The bilateral method and our robust method give quite different index series for clothes. After 10 periods the difference is almost 20 percent in West-Finland and Uusimaa. The same holds also for other NUTS2 regions. Results for shoes are different – methods are quite closely related. This holds also for other NUTS2 regions.

Clothes includes hundreds of detailed commodity groups including commodities which have quite short life cycle. We think that our robust method has recognized the pricing mechanism in the end of the year 2020 (prices have practically not changed), but in the beginning of the year 2021 very heavy covid-19 restrictions declines prices of clothes. In March 2021 the prices begin to increase and almost reaches their initial level.

Index series for shoes is instead quite 'alarming' because our robust method goes almost 'hand-in-hand' with bilateral strategy, meaning that the construction strategy for shoes needs further clarification about strategy, pricing mechanism, more knowledge of additive decomposition of value change (i.e. changes of quantities and prices) and behavior of life cycles of commodities.

5 Conclusion

In this work, we have argued that the index construction methods based on bilateral price-links do not capture sufficiently well the life-cycle effects of the different vintages of clothes and shoes. As a remedy, we propose a simple, robust method, based on conditional average prices of the vintage by age. This method shows a superior performance in correcting the life-cycle bias for clothes, while the results are unclear for shoes. Thus, this work leads to improved quality in the price indices of clothes.

The findings presented herein are of general interest for three main reasons. First, countries are increasingly adopting scanner data and other complete microdata sets for constructing price indices, presenting serious challenges to the current methodologies, which are often not sufficiently confronted with the complexities and heterogeneities found in the real-world data. In order to fully benefit from the big data revolution, the statistical offices need to make sure that the methodologies applied are suitable for these new, more complete, but more complex, data sets. Third, it is likely that there are more product categories like clothes and shoes that require adopting more suitable methods due to life-cycle effects and strong attrition.

Our paper is only the first step towards correctly measuring price changes in complete microdata sets for products with significant life-cycle effects, and strong turnover of vintages. Countries should test this method using similar data sets and consider a wider range of product categories as well, where similar problems are likely to arise. Further research should consider at least certain high-tech categories, such as mobile phones. Regardless, careful testing of the effects of quality changes during the sample is important, so that we do not mask the larger trends of improved efficiency and products.

References:

Eurostat, (2021): ‘HICP recommendations on bridged overlap’.

Hsiao, C. (1986): **‘***Analysis of Panel Data*’, Cambridge University Press.

Suoperä, A. & Auno, V. (2021): https://papers.ssrn.com/ ‘*Hedonic Index Numbers for Rents of Office and Shop Premises in Finland’.*

Suoperä, A., Nieminen, K., Montonen, S. & Markkanen, H. (2020): https://papers.ssrn.com/ ‘*Comparing Basic Averages, Index Numbers and Hedonic Methods as Price Change Statistics*’.

Vartia, Y. & Suoperä, A. (2018): ‘*Contingently biased, permanently biased and excellent index numbers for complete micro data*’, http://www.stat.fi/meta/menetelmakehitystyo/index_en.html