Missing in action: testing alternative imputation methods in price statistics

Patrick Kelly and Matlhatsi Mogalanyane
Statistics South Africa

Meeting of the Ottawa expert group on price statistics
Rio de Janeiro
May 2019

Imputations are a critical component of any statistical computation and bear special importance for the operational aspects of price statistics. Missing prices for seasonal items are a particular challenge. The literature elaborating imputations for price statistics is repetitive on the basics, but light on technical and analytical detail. Focussing on temporarily missing prices, this paper reviews the advice provided to compilers. The performance of different imputation methods for seasonal and non-seasonal products is assessed through the use of a multi-year synthetic dataset. Tests are conducted on the most appropriate level of imputation and the robustness of the standard options presented in the manuals. Time product dummy regression is tested as an additional imputation technique.

1. Introduction

Missing prices are part of the day-to-day reality of any price index compiler. Unless the item[1] is to be replaced, the go-to response is to impute the price. The basic methods of imputation are well itemised in chapter 6 of the draft revised CPI manual (IMF, 2019), which will be of significant assistance to CPI compilers. However, for a topic of such importance to index compilers, there are surprisingly few studies in forums where they may look to for advice - ILO/UNECE or Ottawa group meetings.

This paper aims to provide insight to the most appropriate methods of imputation, not from a theoretical perspective, but through a data-based analytical exercise to test the performance of different methods.

There are several reasons why a price may be missing, and these can be reduced to two types - permanent and temporary. Should a product no longer be stocked in a particular outlet, it is then permanently missing and requires replacement. This paper is not concerned with this scenario and the consequent substitution rules and procedures.

Temporarily missing prices are usually caused by an item not being available in an outlet at the time of price collection due to its being sold out, or because it is only available in certain seasonal months. In both cases, the price collector assumes that the variety will soon be available again. In the case of a stock-out, the item is expected back on the shelf within the next two months. Seasonal products are expected at the beginning of the next in-season period.

2. Reflecting on the aim of imputation

The default pricing method for a CPI is to record the price of an identical product variety in the same outlet each month. This is known as the matched model method. When an item is temporarily unavailable, it creates a gap in the matched series of prices which may create a bias in the elementary index.

The question of why we impute for temporarily missing prices is seldom asked or answered. One reason is to arrive at an index that is as close as possible to what the index would have been if the price was available. From an analytical perspective, this is the key consideration. It follows that the imputed price should be as close as possible to what the missing price would be. Here the medical adage ‘first do not harm’ is relevant. As we do not know what the values would have been, it is critical to avoid any bias to the index through imputation.

A second goal is to ensure a complete matrix of prices without disturbing the change in the index that was calculated without the price. A complete set of prices in the database has a practical aim of having a t-1 period comparator price for an actual price when it is available and recorded in period t.

[1] The terms item and variety are used interchangeably to describe the object for which a price is recorded.

3. Review of methods and literature

The draft version of chapter six of the new consumer price index manual (IMF 2019) sets out the main methods for dealing with temporarily missing prices. These are overall or targeted mean imputation, to carry forward the previous price, or to simply omit the price altogether. For completeness, the methods are briefly described here.

Overall mean uses the average price change of all observed varieties within an elementary index that are matched to the previous month to calculate a price change for the missing varieties. Should no prices be available within an elementary index, the mean of all prices within the next level aggregate should be used. Usually, varieties making up the sample for an elementary aggregate will include a mix of price levels, sizes, quality and outlet types. The overall mean assumes that the price movement of these will be similar. A key feature of overall mean imputation is that 'it does no harm'. By using the average relative of all the matched observations, the index should not be biased up or down. However, this overall movement might have been different had the price of the missing variety been known. The overall mean is the simplest to program and requires no subjective decision making.

Targeted mean imputation uses the average price change of a subset of varieties - for example within the same geographic area or type of outlet. This method includes that traditionally known as a class mean. A key requirement of this method is an adequate sample size of those items used to calculate the imputation.

Carry forward uses the same price as the previous month - meaning no price change. This method should only be used when the compiler has a high certainty that the price has not changed - for example tariffs set annually.

It is, of course, also possible to simply omit a missing price or *do nothing*, which will yield the same aggregate index change in that month as an overall imputation. However, an imputation facilitates the use of the next available price, thus ensuring a complete dataset. Imputing also returns the price series to its correct level. Omitting the price does not have this 'self-correcting' benefit.

Time product dummy (TPD) regression has long been used as a method for quality adjustment when permanently unavailable varieties require substitution (e.g. Diewert et al, 2007). However, its use to impute temporarily missing prices is uncommon and does not appear as an option in the CPI manual. The regression uses a specified array of data with a longer time series than other imputation methods to estimate the price change. The authors expected the TPD to perform well in imputing for seasonal items given its backward reach.

4. Review of CPI expert group papers

A review of papers and presentations from meetings where CPI compilers would look to for advice reveals a surprisingly sparse coverage of discussion on imputation. Two contributions outline the basic methods for imputation (Erdogan, 2011 and Johannessen, 2018) but do not much advance the discussion. Only Roh and Becker-Vermeulen (2013) have attempted to analytically consider the performance of different imputation methods. Their focus is on the use of imputations in the Swiss CPI when substituting permanently missing items. A synthetic dataset was used to assess the relative strength of imputation methods to link in

the introduction of new varieties of clothing - a strongly seasonal product. Imputations are therefore used as a form of quality adjustment.

Two methods were compared in the Roh and Becker-Vermeulen study - the overall mean, which used the average price change of matched varieties within the elementary aggregate, and a class mean, which used the average price change of varieties that had been replaced in that month by direct comparison.

The study showed that using the overall mean resulted in an index lower than had been originally published. The class mean, in contrast, showed either no effect or a slight upward difference from the baseline index. The upward effect is most noticeable in fashion items and in months where significant seasonal change in clothing varieties occurs. They conclude that the optimal class mean imputation is based on clothing items within the same outlet types.

In order to operationalise the class mean imputation in the Swiss CPI, a larger sample was required to avoid bias and the operational procedures for price collectors were made stricter.

5. Methodology

The aim of this exercise is to assess the performance of different imputation methods to estimate prices for temporarily missing varieties. A 25-month dataset was created based on data from the South African CPI. Any missing prices in this dataset were imputed using an overall mean to ensure a complete matrix for analysis.

The dataset comprises 1 751 varieties covering a selection of products that show different pricing behaviours as listed in Table 1. Ten percent of price observations were deleted at random to create a data set resembling that which may realistically face price statisticians.

Table 1. Description of dataset

COICOP group	Product	Pricing behaviour	Number of observations
Bread and cereals	Bread	Stable	758
Fruit	Peaches	Strong seasonal	10
Vegetables	Broccoli	Weak seasonal	57
Milk, cheese and eggs	Eggs	Stable	414
Clothing	Men’s shorts	Strong seasonal	197
Clothing	Men’s jeans	Stable	180
Furniture	Bedroom suites	Sticky	135

Four imputation methods were applied to the data set to impute the missing prices.

- The overall mean used all available price relatives for that particular product to derive an imputed price.
- The targeted mean used a subset of varieties for that product - which in this case was a specific geographic area.
- The carry forward method uses the price in t-1 as the price in t.
- The TPD imputations were calculated using all data from the product group for the current and previous 12 months.

6. Empirical Results

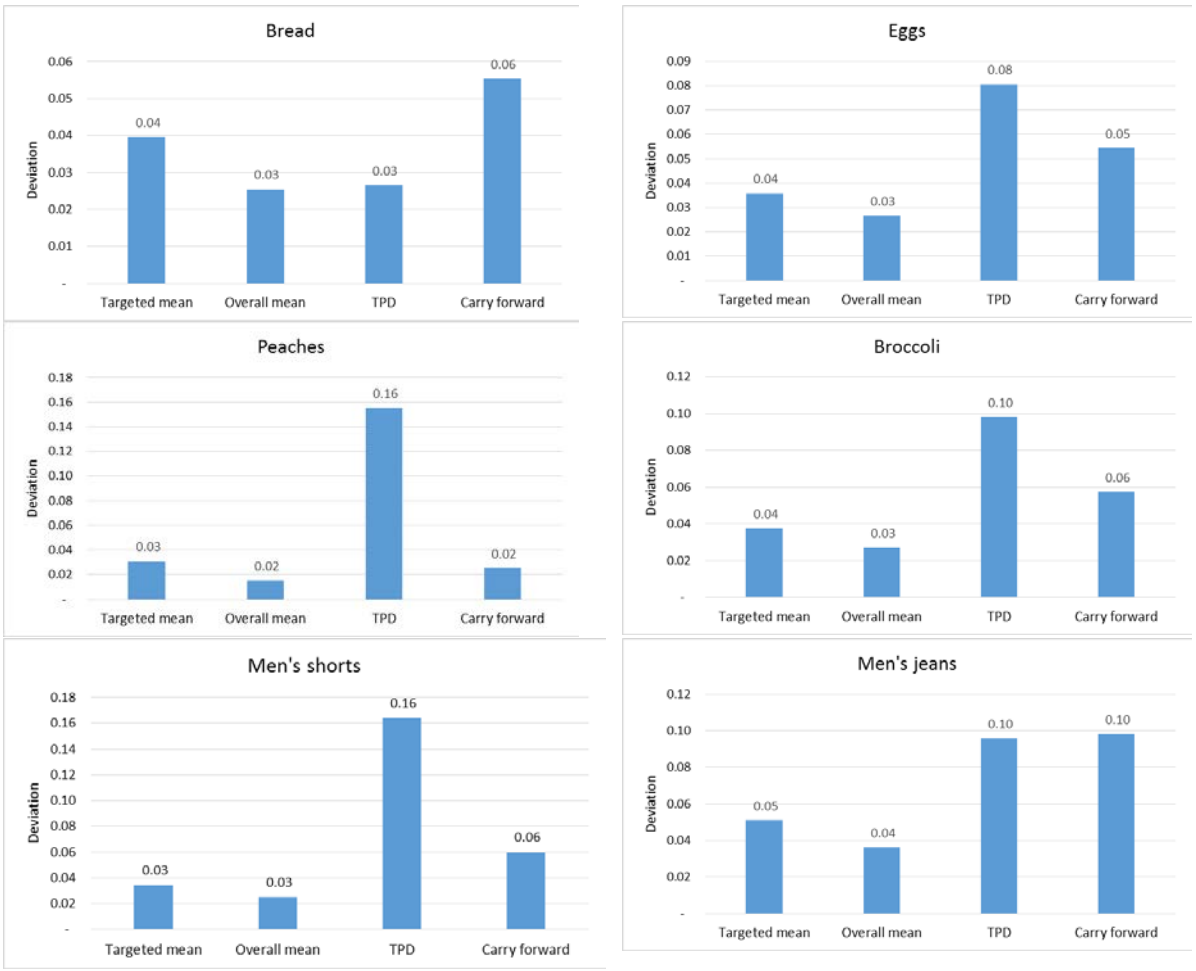
Results are examined at the price and index level. The imputed price is compared with the price deleted from the complete dataset and a mean absolute average deviation calculated. This deviation is represented as a proportion of the original price.

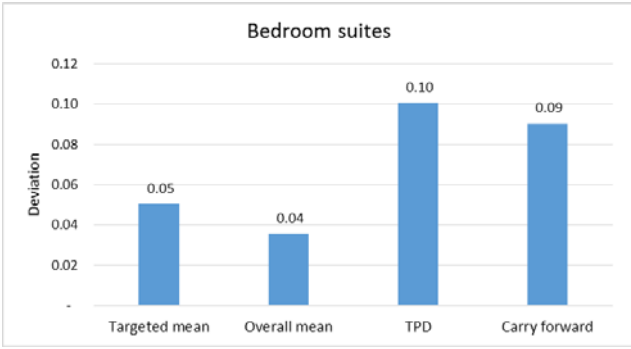
Elementary indices are calculated using the geomean of the price relatives (including those imputed). The average monthly deviation of each imputed index from the original dataset is computed.

a) Price level

Generally the deviation of the imputed prices is low. However, variation is noted between the methods. Prices imputed using the overall mean are closest to the original price for all seven products. In six cases, the targeted mean showed the second lowest deviation from the original price. In all but two instances, the TPD shows the worst result, registering deviation levels several times that obtained by the overall mean. In these two cases, the carry forward method recorded the highest deviation.

Figure 1: Difference between actual and imputed price

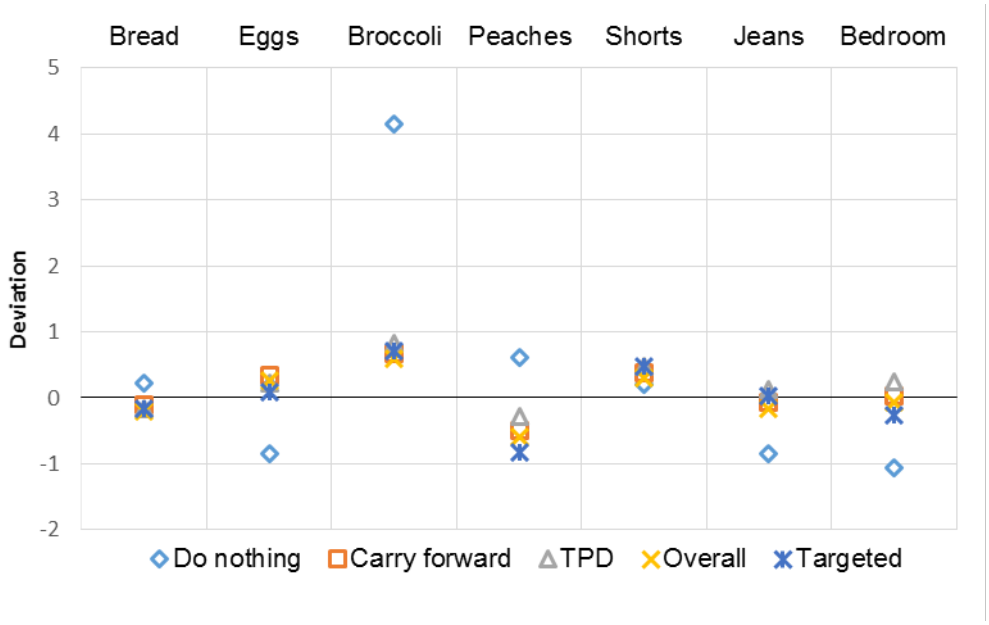




b) Index level

Indices were calculated for a 12-month period for each imputation method, the original data set, and the dataset with no imputations. Figure 2 shows the average deviation of each index from the original index. The graph also indicates whether the imputation caused a downward or upward effect.

Figure 2. Average deviation of index by imputation type



Omitting the price altogether - do nothing - showed the largest deviation in 6 of the 7 products. In four of the seven indices, omitting the price had an upward effect on the index.

All four imputed indices showed similar, small scales of deviation from the original index. In five of the seven products the direction of the deviation caused by the four methods was the same. The TPD imputation showed the second worst level of deviation in aggregate. Although it was expected to pick up seasonal trends, there was no clear evidence of this. Nevertheless, TPD registered the lowest deviation for peaches - a seasonal product - and second lowest for eggs.

Carry forward imputations performed the worst of the three traditional imputation methods. The exceptions were men's shorts and bedroom suites (which has sticky prices) where it showed the smallest deviation.

The overall and targeted mean imputed indices showed similar and small levels of deviation. The targeted mean has a slightly lower deviation in aggregate, but the overall mean is lower in four of the seven products. It is notable that the overall mean performs better for products with some seasonal patterns (peaches, broccoli, men's shorts).

7. Conclusion and recommendations

The study confirms that the overall and targeted mean imputation methods recommended by the CPI manual and generally used in statistics offices are the most reliable. These methods most often had the lowest deviation in the imputed price and the overall index. They meet both the aims stated above of most closely approximating the missing price and of not biasing the index.

The overall mean has the advantage of a larger pool of observations from which to impute compared with the targeted mean. Compilers need to ensure that the subsample for targeted imputations is adequate and properly represents the characteristics of the missing values.

The comprehensiveness of the overall mean may account for why it does better at imputing prices with seasonal behaviour, as volatile prices are moderated. However, this is not true for the TPD approach. Due to its using data for a longer time period, it was hoped that TPD would perform better at imputing seasonal prices, but the study was not able to support this hypothesis. This is an area for further investigation.

Although the results for carry forward were positive for bedroom suites and which have sticky prices, compilers should still be hesitant before using this method. Instituting a carry forward imputation will miss or lag actual price changes when they do happen.

Finally, the study confirms the warning in the manual that it is better to impute a price than to omit the price altogether. The 'do nothing' approach does not meet the aim of 'do no harm' as the cumulative impact of the non 'self-correcting' characteristic of missing prices on the index clearly creates a bias over time.

References:

Diewert, E, Heravi, S and Silver, M, 2007, Hedonic Imputation versus Time Dummy Hedonic Indexes, International Monetary fund working paper WP/07/234. Accessed at www.imf.org/en/Publications/WP/Issues/2016/12/31/Hedonic-Imputation-versus-Time-Dummy-Hedonic-Indexes-21370

Erdogin, C, Practical Experiences with Calculating Elementary Indices and Treatment of Missing Prices, Presentation at the Workshop on Challenges in Consumer Price Indices, 2011, Turkey. Accessed at www.unece.org/index.php?id=22270

International Monetary Fund, 2019, Update of the Consumer Price Index manual. Accessed at www.imf.org/en/Data/Statistics/cpi-manual

Johannessen, R, 2018. Missing items. Presentation at the meeting of the group of experts on consumer price indices, Switzerland. Accessed at www.unece.org/index.php?id=46772

Roh, C and Bekker-Vermeulen, C, 2013. Clothing: the use of class mean imputation in the Swiss Consumer price index (CPI) – analysis and impact on the results. Paper presented at the meeting of the Ottawa Group, Denmark. Accessed at www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Meeting+13+-+Copenhagen,+2013