Imputing Web Scraped Prices

Matthew Mayhew

23 May 2016

Summary

Missing prices cause a problem for price indices made from web scraped data, therefore it is necessary to find a way to deal with them effectively. Imputation is a method of dealing with missing prices, but there are many different techniques to choose from. Following a simulation study it was found that carrying forward the previous price is the best method with respect to minimising imputation bias. There are two effects of imputation on the GEKSJ index that is created from the web scraped; either there is little difference in the indices, or imputation smoothes out volatility and spikes caused by the missing prices.

1 Introduction

The ONS has been producing experimental price indices based on daily price information scrapped from supermarkets websites, thus producing more frequent indices than the traditional CPI collection methods. Sometimes prices cannot be collected, for reasons such as the product goes out of stock, which occurs as well in traditional CPI production, or the scraper breaks. These missing prices have effect on the indices in that they cannot be produced properly. There are two options in solving this, either remove the item from the sample for the day the data is available for then calculate the index, known as a matched sample, or impute the missing price. This report will focus on the imputation option for dealing with missing prices and explore a variety of different imputation methods, assessing their impact on price indices and provide recommendations.

2 Imputation Methods

There are many different types of imputation methods, and out of these methods only three types were tested, which are as follows

1. Carry forward the previous price i.e. $\hat{p}_i^t = p^{t-1}$
2. Class Mean by store or by item type, using:

   (a) Arithmetic Mean

   $$\hat{p}_i^t = \frac{1}{n-1} \sum_{\substack{j \in C \\ j \neq i}} p_j^t$$

   (b) Geometric Mean

   $$\hat{p}_i^t = \left(\prod_{\substack{j \in C \\ j \neq i}} p_j^t \right)^{\frac{1}{n-1}}$$

   (c) Harmonic Mean

   $$\hat{p}_i^t = \frac{n-1}{\sum\limits_{\substack{j \in C \\ j \neq i}} \frac{1}{p_j^t}}$$

   where C is class, for example item or shop.

3. Ratio Imputation: apply average growths of the rest of the items then multiply this by the previous price, using:

   (a) Arithmetic Mean

$$\hat{p}_i^t = p_i^{t-1} \times \frac{\sum\limits_{\substack{j \in I \\ j \neq i}} \frac{p_j^t}{p_j^{t-1}}}{n-1}$$

   (b) Geometric Mean

$$\hat{p}_i^t = p_i^{t-1} \times \left(\prod_{\substack{j \in I \\ j \neq i}} \frac{p_j^t}{p_j^{t-1}} \right)^{\frac{1}{n-1}}$$

   (c) Harmonic Mean

$$\hat{p}_i^t = p_i^{t-1} \times \frac{n-1}{\sum\limits_{\substack{j \in I \\ j \neq i}} \frac{p_j^{t-1}}{p_j^t}}$$

One drawback of imputation is that it may introduce some imputation bias into the result, the imputation bias for price i at time t for is calculated by

$$B_i^t = p_i^t - \hat{p}_i^t.$$

An imputation bias of £0.10 in a price that is £0.50 has more influence than the same imputation bias in a price that is £50, therefore the relative imputation bias is calculated, this is as follows

$$RB_i^t = \frac{B_i^t}{p_i^t} = \frac{p_i^t - \hat{p}_i^t}{p_i^t}.$$

This Relative Imputation Bias is then used to determine the optimal imputation method. For the example given above the Relative imputation bias for the £0.50 item is 0.2 and for the £50 item it is 0.002, therefore imputation has more influence on any indices created which involves the first price than the second. The direction of the imputation bias is also important as if the imputation bias over a group of items is consistently is negative, i.e. imputed prices are larger than collected prices, then an index produced from this price may possibly be higher from using the imputed data, and the converse will happen if there is a positve imputation bias on the price. The aim is to find an imputation method which minimises the relative imputation biases, and is therefore our best estimate for a missing price. The Absolute Relative Imputation Bias also needs to be check; this is taking the absolute value of the Relative Imputation Bias

3 Simulation Study

To find the method that minimises the relative imputation biases, the following method was used:

1. Find an area of the web scraped data with no missing prices.
2. Remove a sample of the prices.
3. Impute the prices.
4. Calculate the average relative imputation biases.

Two data sets were created, from time periods which did not have any missing prices in it. These time periods are the first three weeks of that the data was collected in, 01/06/2014-22/06/2014, and the second data set is four weeks from the middle of the scraping period, 12/02/2015-12/03/2015. Dataset 1 has 3989 products, and dataset 2 has 3599 products. As the datasets are composed of approximately 100000 prices it was decide that a 10% sample of 10000 prices would be removed. The number of prices to remove for each item and shop pair strata was calculated using a proportional allocation method, which replicated the pattern of missingness in the underlying data. This makes sense because items with larger number of prices had more products available to buy, which may go out of stock quicker due stores would holding less in their inventories in order to have more of a range of products. After the imputation was performed, the relative imputation bias of the imputation was calculated.

Then two means were taken, the first $\overline{\mid RB \mid}$, the mean absolute relative imputation bias, and the second $\overline{RB}$, the mean relative imputation bias. These two means were calculated for each of the imputation methods and for the prices, the Chained Daily and the GEKS index. Figure 1 shows the mean absolute relative imputation bias for each imputation method on both datasets. The imputation method that minimises the mean absolute relative imputation bias, on the prices and on the indices, is carrying forward the prices. The second best imputation method depends on whether on index formula, for the Chained Daily Index it is the average growth methods, this is also the second best for the prices, whereas for the GEKs index the second best set method is class mean imputation, though the best class depended on the time period. However, the direction of this imputation bias will affect the growth rates of the index as well, so a look at the direction of the imputation bias, through the use of mean relative imputation bias, would aid a better decision on the choice of method. Figure 2 shows this.

The same results hold for the mean relative imputation bias as for the mean absolute relative imputation bias, though the magnitude of the relative imputation bias confirm that carrying forward may not affect the growth rate of the index as the rounded value would be the same.

Figure 1: *Mean Absolute Relative imputation bias*

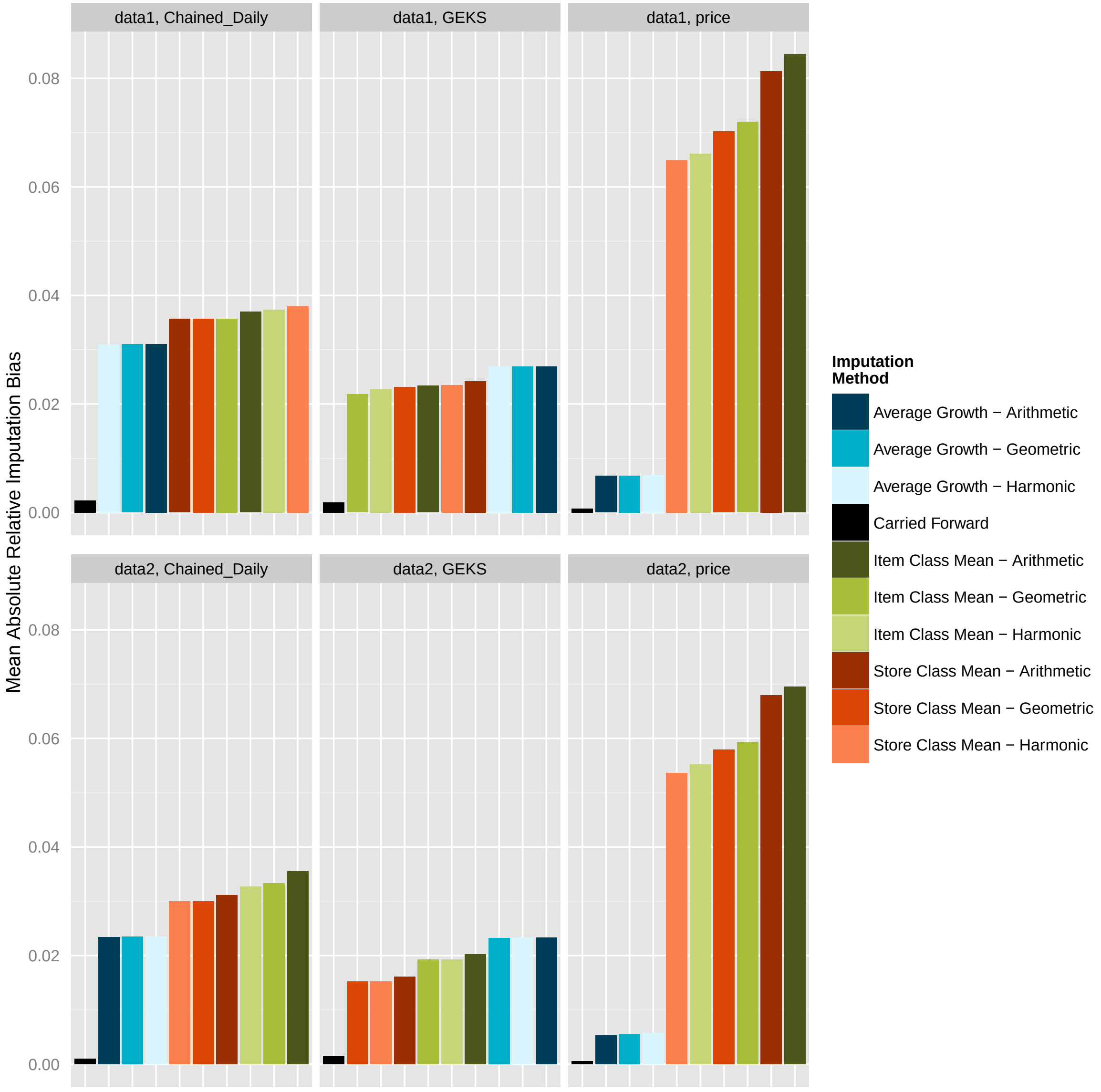
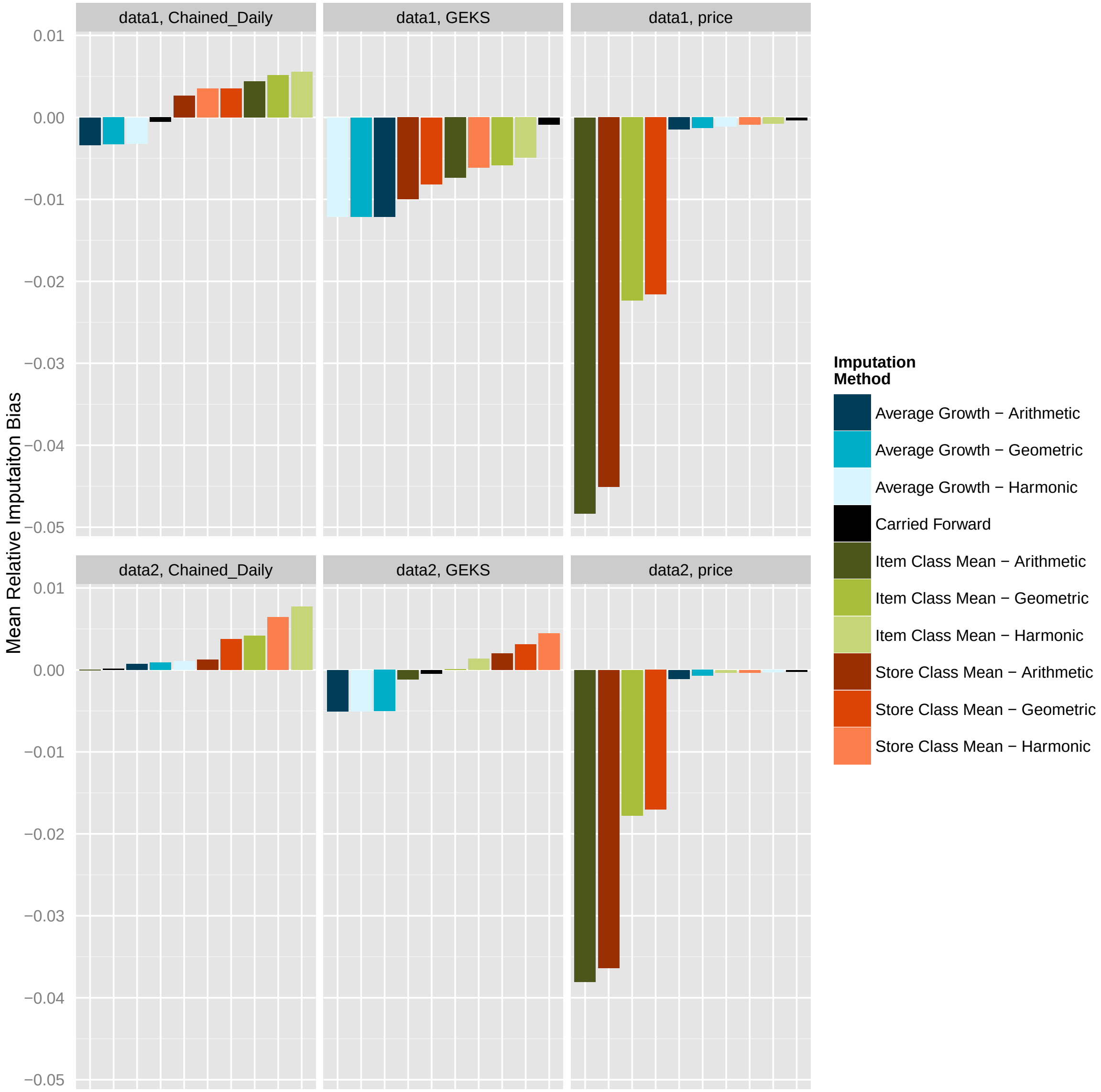


Figure 2: *Mean Relative imputation bias*



4 Justification for carrying forward prices

Figure 3 shows the distribution of the average time between price changes in the web scraped dataset. The average time between price changes has been calculated as total number of daily price quotes / number of price changes. The figures exclude any item that appears on the dataset for less than 31 days.

Figure 3: *Distribution of average time between price changes, all items, raw data June 2014 to February 2016. Median 120 days (light blue) arithmetic mean 181 days (green).*

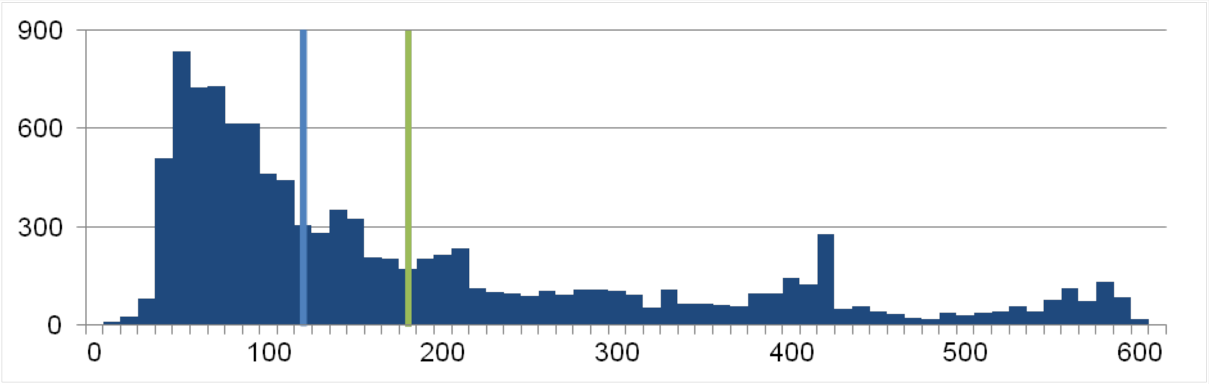


Figure 3 shows that the majority of prices do not change regularly, in fact many do not change at all across the dataset. This supports the recommendation to carry forward previous prices.

5 Recommendations

With the optimal imputation techniques found with respect to different objective functions some recommendations are to be made depending on whether or not the web scraped prices are to be used to supplement the CPI in the future, as there are rules from Eurostat and the ILO that the CPI has to follow. Table 1 shows these recommendations for imputations not in the base period.

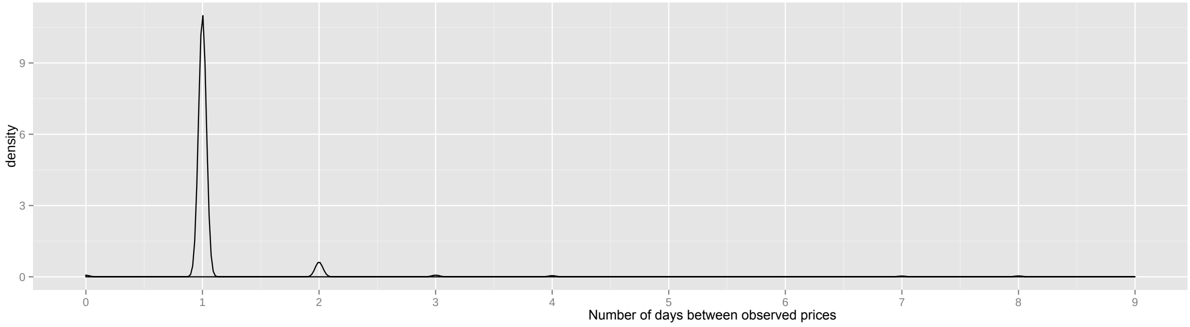
Table 1: *Recommendations for Price Imputation*

Impute on	Data used to supplement CPI	Use in Experimental Statistics only
Prices	Geometric Average Growth	Cary Forward
Chained Daily	Geometric Average Growth	Cary Forward
GEKS	Geometric Class Mean by Store	Cary Forward

6 How long to impute for?

Imputing prices is a good way to deal with missing prices so that a more consistent sample size is kept over all the period of interest, but sometimes a product may go out of stock for a significant period of time and either get reintroduced or removed from the market all together. Therefore, it may be unwise to continually impute the prices in either of the situations, as it would introduce stability into the index or the index would not be representative of actually price movements. To decide the appropriate number of days over which to impute the prices, the number of days between observed prices was calculated and a Gaussian kernel density estimator of the distribution was calculated for all of the items together and for each item individually. Figure 4 shows the KDE for all items in the cleaned dataset and 5 shows the KDE for each item individually.
Observing the distributions in figures 4 and 5 there is a spike at date difference of 1 and then a second one at 2 then a small one at 3. A difference of one means that the prices are observed on contiguous

Figure 4: *KDE date difference for all items*



days. After removing data where observed price dates were contiguous the average number of days between price observations was 2.7 days so it would be recommended that prices should be imputed for is 3 days then an item be removed. If there is a scraper break larger than 3 continue imputation until scraper is fixed unless this is larger than a week then stop imputation. Figure 6 shows that the values of 3 and 7 days are not arbitrary, as the percentage of products that have missing prices is 89% and for 7 days is 99%, therefore imputing for a week does cover almost all of the missingness.

Figure 5: *KDE date difference for different items*

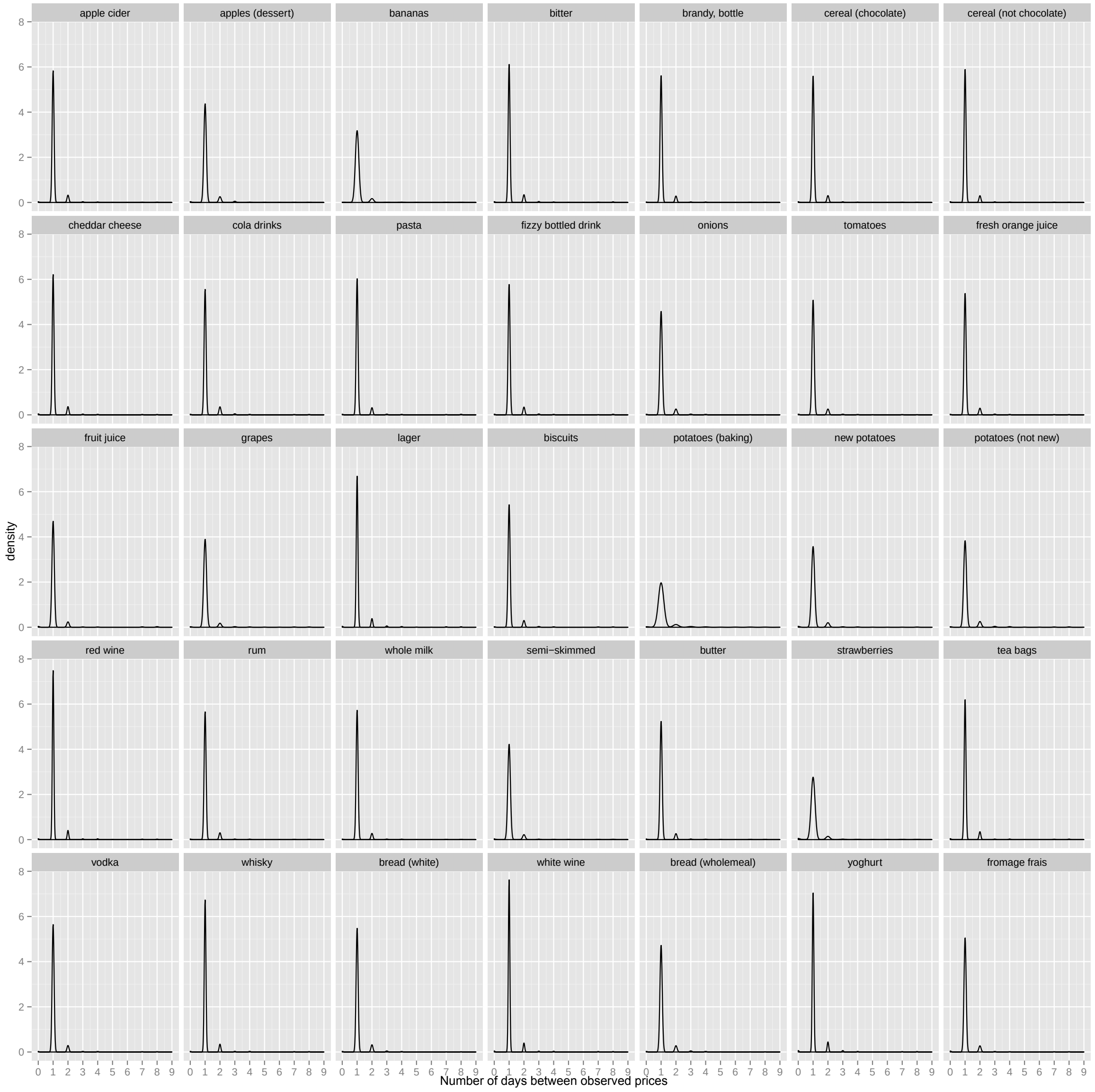
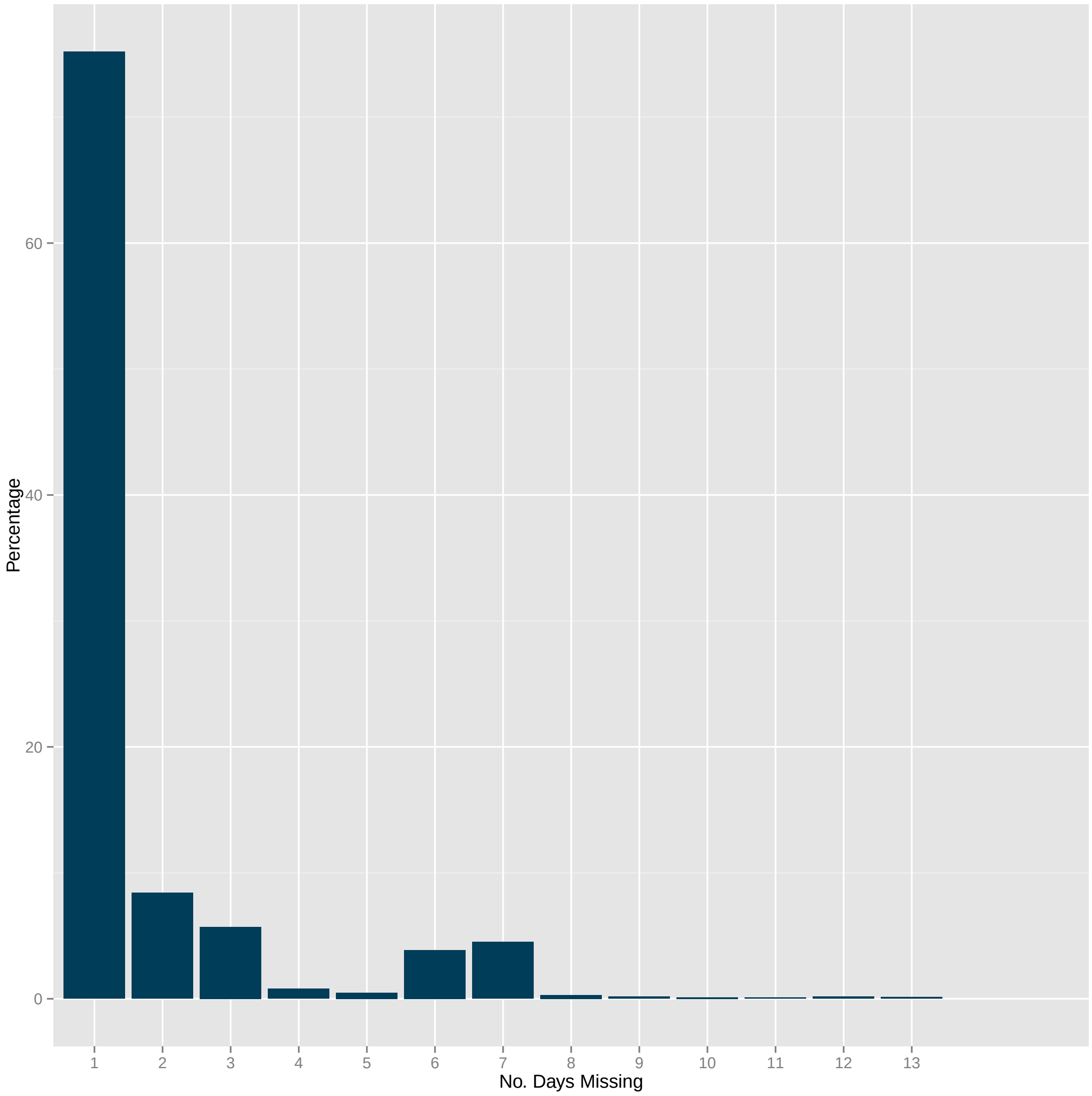


Figure 6: *Percentage of products by number of days of missing prices*



The 7 day imputation rule for scraper breaks is also justified by looking at the number of days for which a scraper break occurs. Table 2. Here the majority of scraper breaks are less than a week, imputing for a maximum of seven days would stop irregularities in the index series caused by missing prices.

Table 2: *Length of scraper breaks by supermarket, June 2014 – April 2016*

Break Length (Days)	Supermarket			
	Sainsbury's	Tesco	Waitrose	Lab Failure
1	22	15	16	12
2	1	2	1	1
3	1	2	2	2[1]
4	1	0	0	0
6	1	1	0	0
7	1	1	0	0
26	1	0	0	0
34	1	1	1	1

7 Does imputation have an effect on the indices?

For the purpose of this section only the GEKSJ index has been considered, as an indication of the impact of the imputation, when imputation was done over the whole of the period of collection. Looking at the results there are two different types of effects imputation has, these effects are either:

1. Indices calculated using imputed data are almost identical to indices calculated using non-imputed data.

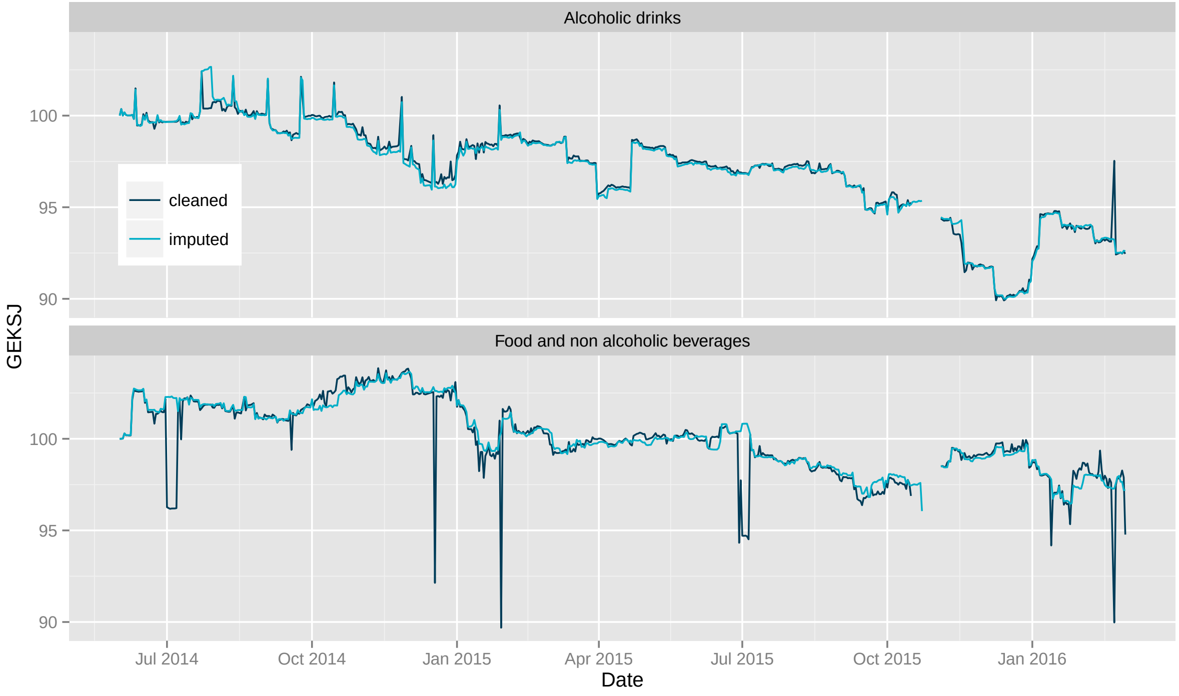
2. Indices calculated using imputed removed irregularities and smoothes out the series

Figure 7 shows both cases.[2] For the Alcoholic Drinks, the imputed GEKSJ index is close to the cleaned GEKSJ index therefore imputation doesn't change the index. On the other hand, the food and non alcoholic beverages indices shows the second case, as every so often there is a sudden spike in the non-imputed indices. This is because the food index is an aggregate of the lower level indices using expenditure weights from Living Costs and Food Survey, these weights sum to one, so these jumps are due to a missing price causing missing indices and therefore the weights wouldn't sum to one. The imputation of prices causes the indices to exist then when aggregating the weights do sum to one and therefore the growth in the index is purely down to a change in prices not a change in the weights. From the imputation a better understanding of inflation that consumers experience, as when the scraper breaks the consumer will still be able to buy the products from the websites. For cases when the product goes temporarily out of stock, a consumer in a different part of the country may still be able to buy the product as the supermarkets scraped are national chains, and do change the products available on the website depending on the products which are available in the stores in the locality of the consumers address.

[1] This number is greater than the number of three day breaks in the Sainsbury's web scraper as the three day lab break is part of a longer break in the Sainsbury's web scraper.

[2] The Break in the Series is caused by a larger scraper break and due to the imputation rules there is still missing data

Figure 7: *GEKSJ for the Food, and Alcoholic Drinks COICOP divisions*



8 Conclusion

In conclusion, using imputation is a good method to deal with missing prices due to stock unavailability and scraper breaks. This is because it has a favourable effect on the indices in that it stops some volatility and level changes due to changes in the weights. The best imputation method was to carry forward the prices as it had the lowest average relative imputation bias. Following this work imputation was carried out on the prices used in an update of the research into using web scraped data in price indices.

Acknowledgements Thanks to contributions from Iolo Tomlinson, Joseph Winton and Tom Hunter Smith for their contributions and expertise in carrying out the simulation study.