A Simulation Study of Multiple Imputation Methods for the Producer Price Index

Yoel Izsak, Monica Moleres
Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

Abstract

The Producer Price Index at the Bureau of Labor Statistics currently uses cell mean imputation for missing price data. In the time since the implementation of the current process, multiple imputation methods have become much easier to use on large data sets. In this study, we investigate alternatives to the current procedure. We examined a few different multiple imputation methods with packages in R, including: CART, Random Forest, and AMELIA (bootstrap EM algorithm). We also introduce a hybrid imputation method combining both cell mean and random forest techniques. Success of imputation for the missing prices was measured by RMSE (Root Mean Squared Error) of PPI Index estimates. Results from the study will be discussed.

Keywords: imputation; model comparison; ppi; government statistics

1. Introduction

The Producer Price Index at the Bureau of Labor Statistics currently uses cell mean imputation for missing price data. In the time since the implementation of the current process, multiple imputation methods have become much easier to use on large data sets. In this study, we investigate some multiple imputation algorithms as alternatives to the current procedure.

Section 2 of this paper provides a summary of the PPI index estimation. In Section 3, we discuss the various multiple imputation algorithms that we studied: Classification and Regression Trees (CART), Random Forest, AMELIA, and Predictive Mean Matching (PMM). Section 4 details the setup of our simulation study. In Section 5, we present the results of using multiple imputation algorithms on PPI data. Section 6 details a hybrid imputation algorithm which uses both cell mean and random forest. And Section 7 investigates some different parameter values for that hybrid algorithm.

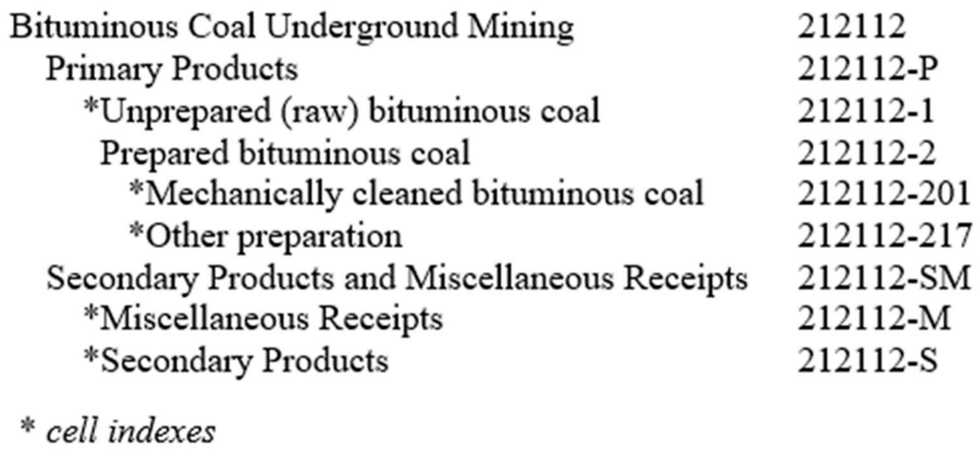
2. Overview of PPI

The Producer Price Index (PPI) is a key economic indicator that measures the average change over time in selling prices received by domestic producers of goods and services. It serves as one of the nation's inflationary indicators, particularly for the business sector of the economy. The monthly average price change is made up of over 50,000 price quotes per month in the form of item prices from sampled establishments. Items are grouped into cells according to their product similarity, or industry. Items usually have prices reported monthly.

The PPI publishes three primary outputs. These are industry indexes, commodity indexes, and Final Demand – Intermediate Demand indexes. The industry structure upon which the PPI is based is the North American Industry Classification System (NAICS). Under

NAICS, establishments that use the same or similar processes to produce goods and services are grouped together. Each month the PPI publishes indexes at the 6-digit NAICS level along with more detailed product level indexes. In addition, the PPI produces detailed commodity indexes. The data collected by industry are regrouped into commodity classifications without regard to the particular industry in which they are produced. The PPI's Final Demand – Intermediate Demand (FD-ID) indexes are aggregations of the commodity indexes.

Figure 1: Example of Index Aggregation Within the Coal Mining Industry



Bituminous Coal Underground Mining	212112
Primary Products	212112-P
*Unprepared (raw) bituminous coal	212112-1
Prepared bituminous coal	212112-2
*Mechanically cleaned bituminous coal	212112-201
*Other preparation	212112-217
Secondary Products and Miscellaneous Receipts	212112-SM
*Miscellaneous Receipts	212112-M
*Secondary Products	212112-S

** cell indexes*

The index for each industry or commodity group has a defined aggregation structure. This structure includes the detailed product cells to which items are assigned and the order of aggregation. The lowest level cells are combined to form higher-level aggregate cells.

Index Estimates are calculated using a modified Laspeyres index formula. The modification differs from the conventional Laspeyres in that the PPI uses a chained index instead of a fixed-base index. Chaining involves multiplying an index (or long term relative) by a short term relative. This is useful since the product mix available for calculating indexes of price change can change over time. (Sheidu)

Establishments that elect to participate in the PPI survey enter item prices monthly. If a company does not report a price for a particular month, that price must be imputed. In the PPI the primary method of imputing for missing prices is the cell mean method. This process assigns the average price change for all reported prices in the cell to the missing items. One benefit of this method is that the index value for the cell is the same, whether the items with missing prices are included or not.

3. Methods Evaluated

The Producer Price Index at BLS currently uses standard cell mean imputation to deal with missing item prices. In the years since this was implemented, multiple imputation has become widely accepted. For this study, we compared various multiple imputation methods to our current cell mean imputation in an attempt to improve accuracy for our index estimates. Multiple imputation methods create numerous imputed data sets, and then use an average (or some other calculation) of the various imputed values to create a final imputed point estimate. We conducted a literature review to look for multiple imputation methods that would both fit our data and be computationally feasible. We used several packages in R, designed for multiple imputation. The goal of this study is to determine the best imputation algorithm for PPI index estimation. We did not consider variance or the distribution of prices as a factor in deciding which method to use. Our measure for comparison between methods is the Root Mean Square Error of the index estimates.

The procedure for multiple imputation (as outlined by Rubin, 1987) is a series of steps:

1. Fit the data to an appropriate model.
2. Estimate the missing data using the selected model.
3. Repeat steps 1 and 2 multiple times for each missing data point.
4. Perform data analysis.
5. Average the values of the parameter estimates, obtained from each model to give a single point estimate.

In this paper, we evaluate four multiple imputation methods, along with a hybrid method that we developed[1], against the current cell mean imputation. Our variable of interest Y that we are attempting to impute is item price. However, some items will have missing values in X (predictor) variables also. These missing X values are treated differently, depending on the algorithm. Most of our imputation schemes are run from the MICE (Multivariate Imputation by Chained Equations) package in R, which imputes missing values for X variables before imputing Y.

From https://cran.r-project.org/web/packages/mice/mice.pdf

> *"The mice package implements a method to deal with missing data. The package creates multiple imputations (replacement values) for multivariate missing data. The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model."*

Our tested imputation algorithms are:

• CART (MICE): This method performs imputation using Classification and Regression trees. CART is composed of a decision tree, created by recursive partitioning, where each fork is a split in a predictor variable and each node at the end has a prediction for the target variable. The CART algorithm is also the basis for Random Forest.

• Random Forest (MICE): Random Forest is an ensemble method, where each model is made up of a large number of decision trees. The random forest model combines the predictions of all decision trees for a more accurate prediction. The ensemble method helps to compensate for overfitting in some trees. MICE uses the Random Forest algorithm by Breiman (2002).

[1] The hybrid method is detailed in Section 6.

- Predictive Mean Matching (MICE): PMM is a type of hot-deck method, where donors are found based on the predictor variables being used in the model.

- AMELIA: AMELIA also creates multiple imputed output data sets. It bootstraps the incomplete data, and then uses the EM algorithm to impute missing values.

4. Simulation Setup

The first step in any missing data problem is to categorize the missing data as: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Multiple imputation is implemented in most statistical software under the MAR assumption. Data is considered MAR if the instances of missing data are related to an observed variable in the dataset. We determined our data to be MAR, as an establishment's major producer status in its industry affects the probability of non-response.

Except for a few industries, we do not have access to prices for our missing items, even well after our publication dates. This means we can't test our imputed prices against the real values. So, we created multiple simulated data sets for each month tested.

We used the following process to simulate data sets for testing:

1. Create a data set D_m, containing all sampled items in a given month m. For this study, m = [Feb2019 – Dec2019, Mar2020 – May2020].
2. We define our population as P_m, which includes all items in D_m with a reported price.
    a. We calculate index estimates (our estimate of choice to check --- change this) using the values from P_m. This is our control to test the imputations against.
3. We then create a simulated data set S_m, artificially setting some prices in P_m to be missing.
    a. Average item nonresponse for PPI is 40%. We use simple random sampling to select 40% of items in P_m as missing, with appropriate proportions based on major producer status.
4. Multiple simulated data sets S_{m1}, S_{m2}, …. were created for each month of data. On average we tested 5 simulated data sets in each month, over approximately a year of PPI data.
5. The missing prices in each S_{m1}, S_{m2},… are then imputed using the various methods in Section 2. We calculate index estimates for each imputed data set. These index estimates are then compared to the index estimates from 2a.
6. RMSE (Root Mean Squared Error) is calculated for every price index in PPI, comparing the difference between each imputed data set and its control data set. This gives us an estimate of how close the imputation can get us to our "true" index estimates.

5. Comparison of Multiple Imputation Methods

5.1 Variable Selection

The PPI database includes approximately 200 variables regarding item and survey unit characteristics. This list of possible X variables was narrowed down at first, in three ways.

1. Exclude variables with no possibility of predicting price.
   - *Example*: Company Name

2. Exclude variables that are duplicates, or very similar to other variables.
   - *Example*: There are variables for Street Address, City, State, etc. We excluded all geographic variables, other than State and Region.

3. Certain variables are not feasible in specific software packages, due to run time.
   - *Example*: Product Code is a categorical variable with over 500 different products. This caused some software packages to crash.

All remaining variables were then tested in the various R packages outlined in Section 2. RMSE of index estimates was used to determine the optimal predictors for each algorithm. There are six total predictor variables that were found useful among the various imputation algorithms. Variables used in each algorithm are below in Table 1.

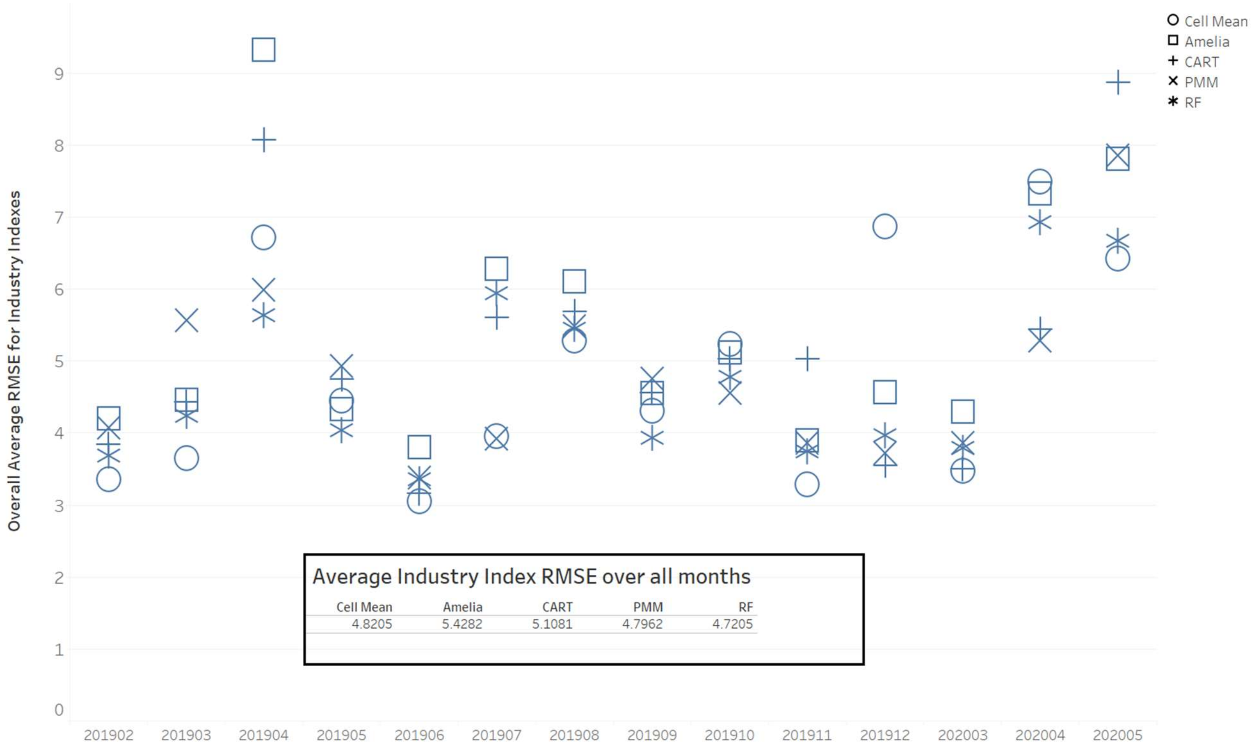
Table 1: Predictor Variables used in Each Imputation Method

	CART	Random Forest	AMELIA	PMM
Product Code	X			X
Major Producer Flag	X	X	X	
Industry code	X	X		
Region	X	X	X	
Multi Hit Item in Sampling	X	X		
Item Weight	X	X	X	X

5.2 RMSE Comparison of Methods

The comparison metric in Figure 2 is RMSE averaged over all indexes in PPI. In each month, we calculated an average RMSE for each simulated data set $S_{m1,}S_{m2}$, …., S_{m5} and then averaged those over all data sets in a month. RMSE is an error statistic, so the best methods are those with lower RMSE.

Figure 2: RMSE for Different Multiple Imputation Methods



Average Industry Index RMSE over all months

Cell Mean	Amelia	CART	PMM	RF
4.8205	5.4282	5.1081	4.7962	4.7205

Figure 2 displays results for each model as defined in Table 1. The current Cell Mean performs quite well overall. There are only small improvements in overall RMSE from other methods, although Cell Mean does perform worse than alternative methods in April 2019, December 2019, and April 2020. As the comparison metric is an average of hundreds of indexes, even small differences are notable.

None of the multiple imputation algorithms consistently performed better than Cell Mean on PPI data, so the focus turned towards possible improvements to the current cell mean imputation process.

6. A Proposed Hybrid Imputation Algorithm

There are 2 main issues with the current PPI cell mean imputation process:

1. The only requirement for imputing a cell mean is a single reported price. This can be a problem, especially in larger cells where one reported price may be used to impute a large number of missing prices.
2. If there are no reported prices in a lowest level cell, we then use aggregate index cells to calculate a cell mean for the lowest level. This tends to impute poorly, since aggregate indexes may contain multiple cells that produce different types of items. The current cell minimum of one reported price was originally set to avoid using aggregate indexes wherever possible in imputation.

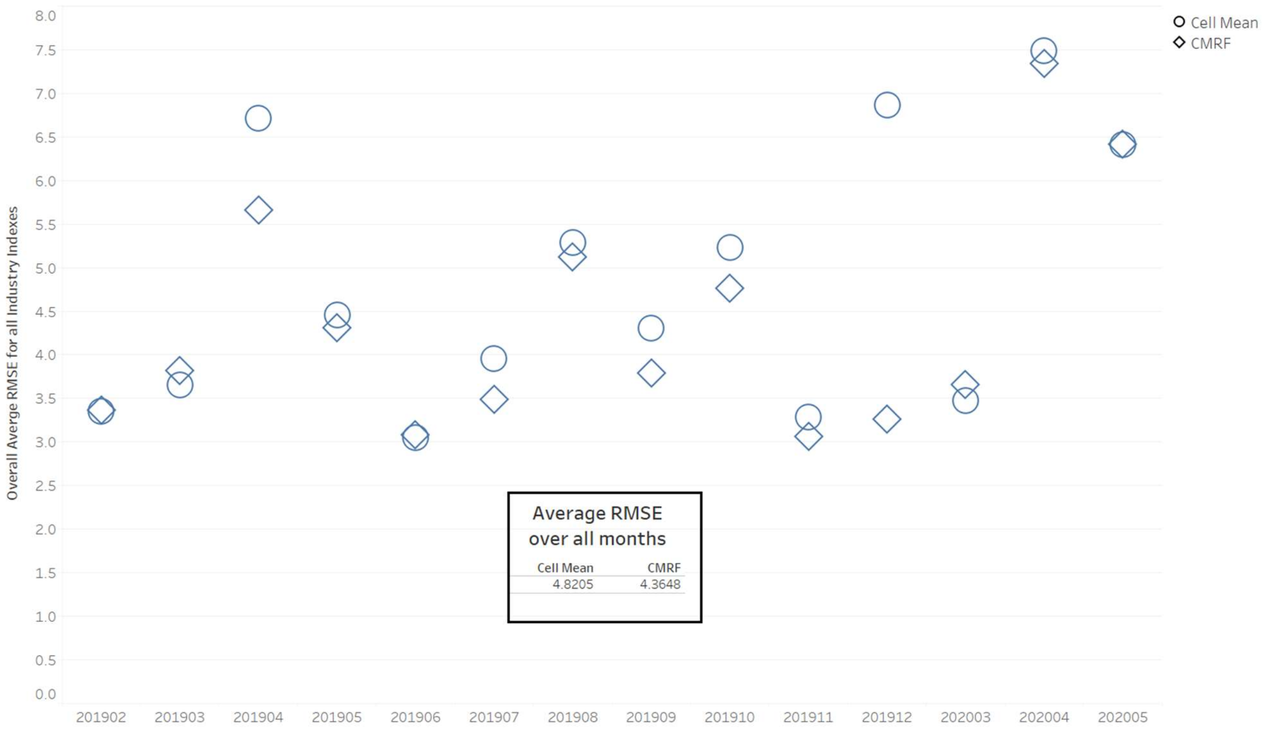
In an attempt to create an improved version of the current process, we tried the following process:

1. Use cell mean imputation, with the following minimum requirements: two reported prices; and reported prices must represent at least 25% of the total weight in the cell. If these requirements are not met, proceed to step 2.
2. Any items unable to be imputed via cell mean at the cell level are instead imputed using the Random Forest model from Section 3. We chose Random Forest, because it seems to improve all situations where cell mean does not work well, and the software package has a reasonable run time.

We refer to this method as CMRF in the rest of the report. The 25% good weight threshold and two good price minimums were selected by educated guesses, and we test other minimums later in the report.

Figure 3 shows average RMSE for both cell mean and CMRF over all indexes. We see that CMRF is a modest improvement overall but had more of an impact in April and December 2019, which are both months in which cell mean performed poorly.

Figure 3: RMSE for Cell Mean vs CMRF



Average RMSE over all months

Cell Mean	CMRF
4.8205	4.3648

7. Testing Modifications to the CMRF Algorithm

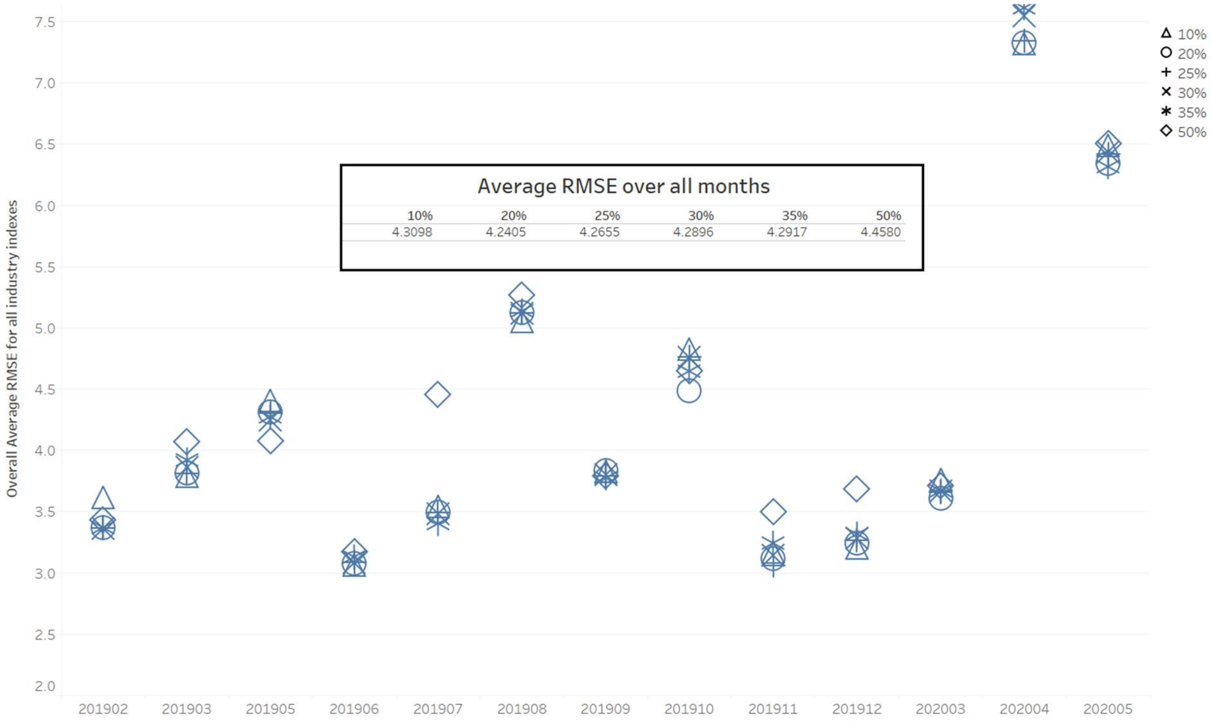
Once we determined that the hybrid CMRF algorithm performs better than our current imputation method, we tested a few different minimums for both thresholds: number of reported prices, and weight of reported prices.

7.1 Reported Price Weight Minimums

We tested a few different reported price weight thresholds for lowest level cells: 10%, 20%, 25%, 30%, 35%, and 50%.

Figure 4 below shows that for most months tested in the simulation study, overall average RMSE is very close for all weight minimums, with a slight increase at 50%. The optimal weight threshold is found to be 20%.

Figure 4: RMSE for Various Reported Price Weight Minimums



Average RMSE over all months

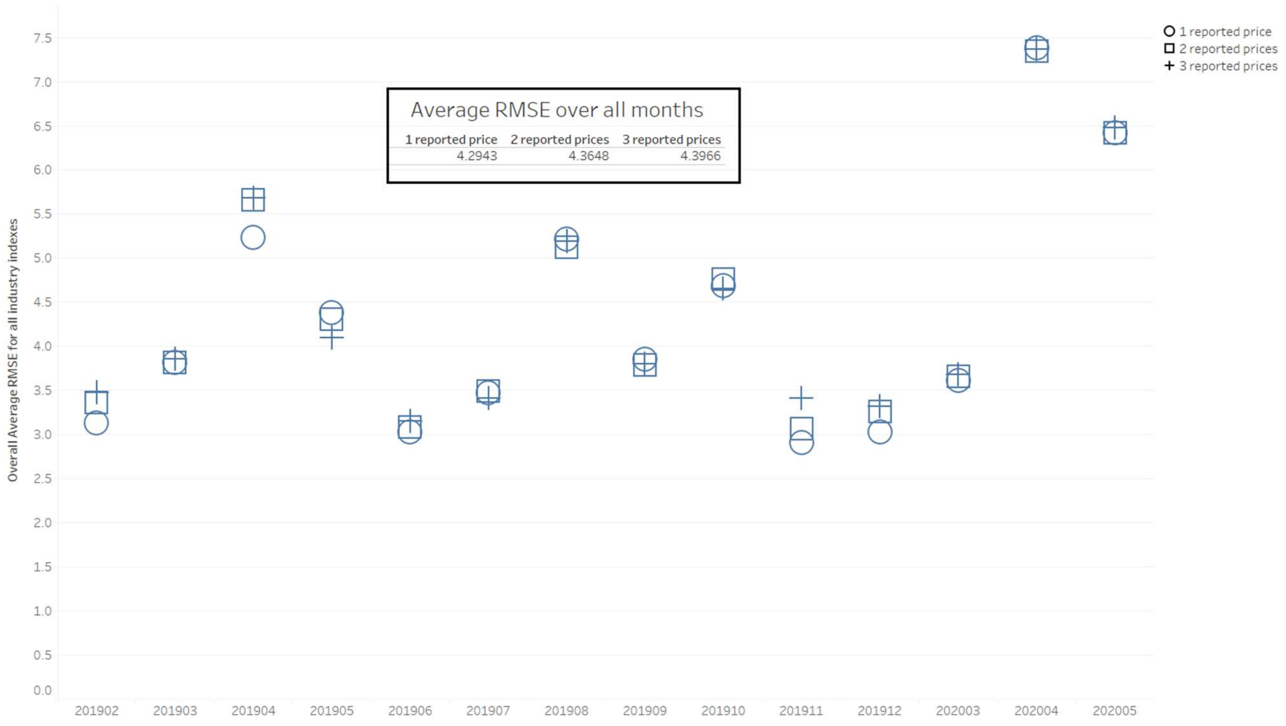
10%	20%	25%	30%	35%	50%
4.3098	4.2405	4.2655	4.2896	4.2917	4.4580

7.2 Minimum Number of Reported Prices

We also tested a few different amounts of reported prices to use as a cell minimum, alongside the 20% reported weight minimum. The current imputation process uses a minimum of one reported price for cell mean imputation in a lowest level cell. We tested the minimum of one reported price against minimums of two and three reported prices.

Figure 5 shows the overall RMSE using these three different reported price minimums as part of the CMRF imputation algorithm. We expect these to give similar results, as most cells with a small number of reported prices would already fail the weight threshold. The RMSEs for all three price minimums are close in most months, with a slight advantage to the minimum of one reported price.

Figure 5: RMSE for Various Reported Price Minimums



Based on these results, we recommended that in the future, PPI should update its imputation method to the CMRF algorithm, using a 20% reported weight minimum.

References

Loh, Eltinge, Cho, and Li (2019), "Classification and Regression Trees and Forests for Incomplete Data from Sample Surveys", Statistica Sinica 29, 431-453.
http://www3.stat.sinica.edu.tw/statistica/oldpdf/A29n122.pdf

Honaker, King, and Blackwell, "Package 'Amelia': A Program for Missing Data".
https://cran.r-project.org/web/packages/Amelia/index.html

van Buuren, "Multivariate Imputation by Chained Equations".
https://cran.r-project.org/web/packages/mice/mice.pdf

Rubin, D.B. (1987), Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc., New York.

Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.

Breiman, L (2002), "Manual On Setting Up, Using, And Understanding Random Forests V3.1".
https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf

Sheidu (2006), "Estimating Missing Prices in Producer Price Index".
https://www.bls.gov/osmr/research-papers/2006/pdf/st060210.pdf