Simulation Study to Compare Imputation at the ELI-PSU Level versus the ITEM-AREA Level

Onimissi M. Sheidu
U.S. Bureau of Labor Statistics
2 Massachusetts Ave. NE, Room 3655, Washington, DC 20212 Sheidu.onimissi@bls.gov

Abstract

The data for the Consumer Price Index (CPI) calculation are regularly collected on a monthly or bimonthly basis. Sometimes, however, not all prices can be collected, which creates missing values in the database used for index calculation. In the CPI, the missing prices are imputed before index calculation. Prior to the creation of the new estimation system, CPI imputed at the Item Stratum-Index Area level. Since the deployment of the new estimation system in January 2015, imputation of missing prices has been done at the Elementary level item-Primary sampling unit (ELI-PSU) level. As a result of this transition, the Imputation Research team was asked to analyze the effect of this switch. In this paper, we conduct a simulation study that compares the imputation of missing prices at the ELI-PSU level versus at the Item Stratum-Index Area level. The results of the study show that imputing at the ELI-PSU level is better than imputing at the Item Stratum-Index Area level.

Key words: Consumer Price Index, imputation, generating missing values, missing completely at random, missing at random, not missing at random.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

1. Introduction

1.1 Overview of the CPI Index

The CPI is a monthly index that measures the average change in prices of a market basket of goods and services bought by urban consumers. It is a weighted average that is currently constructed from over 7700 basic indexes, which correspond to 32 geographic areas segmented into 75 primary sampling units (PSUs).[1]

The CPI is made up of two components – Commodities and Services (C&S), which is about 72.5%, and Housing, roughly 27.5%. This study focuses on the C&S component of the CPI. To calculate CPI, all items within each index area are stratified into item strata, which are later combined into higher item and area aggregates. Thus, an item stratum contains one or more uniquely identified and narrowly defined unit of goods and services known as an entry level item (ELI). Items are mapped to outlets within each index PSU.

[1] In January 2018, the CPI changed the area design.

The CPI uses a two-stage sampling design, with each stage conducted with a systematic probability proportional to size (pps) sampling process. Every year, a sampling frame of the item universe is created by combining all four regional item universes for the two most recent years of Consumer Expenditure (CE) data. During a sample rotation period, which occurs twice a year, samples of ELIs are independently selected from each stratum in the sampling frame within each index PSU by a systematic probability proportional to size sampling procedure. ELI weights are derived from the expenditures reported in the CE data. The Telephone Point-of-Purchase Survey (TPOPS) conducted by Census for BLS serves as the source of outlet sample frames and outlet weights.

The TPOPS provides outlet details and dollar amounts spent on the purchases of groups of items known as Point of Purchase Survey (POPS) categories. A POPS category is a group of commodities and services that are usually sold in the same outlet. A POPS category is composed of one or multiple ELIS.

Outlet frames, total daily expenditure estimates, and selection probabilities are derived from TPOPS data for each PSU-POPS category-sample replicate. For the purpose of sampling variance estimation, the sample for each self-representing PSU is divided into two or more independent subsets known as replicates. Each replicate-PSU contains a single subset of independently selected ELIs and outlets for all item strata within the PSU.

Like the ELI selection procedure, outlets are selected via systematic pps from sampling frames of establishments for each PSU-sample replicate for POPS categories corresponding to selected ELIs in item sampling. The selected ELIs are then priced in sample outlets on either a monthly, bimonthly, or seasonal basis. There are some items not included in the TPOPS that are known as non-POPS items. For these items, separate sample designs are constructed.

The CPI is calculated every month, and the data used for the calculations are collected typically on a monthly or a bimonthly basis. The collection of the data is expected to yield complete price data for every item unit; however, in reality the prices of some items are left empty or unreported for some outlets resulting in the problem of missing data values. This problem is resolved in CPI by implementing appropriate imputation procedures during the index computation. See BLS Handbook of Methods for more details.

1.2 General overview of imputation procedure

Imputation is a method of filling in missing values with substitute values to generate a complete data set for standard statistical analysis. Imputation can either be single or multiple. Single imputation is the process of filling in missing values just once. Multiple imputation means conducting single imputation multiple times and using the average of the estimates derived to fill in for the missing value. Any imputation method is based on the assumption that the distribution of the missing values is similar to the observed data. The goal of any imputation is to minimize data bias created by missing values and not to predict the missing values.

The first step in the imputation process is to evaluate the possible causes of the missing values. According to Rubin (1976), there are three types of missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). MCAR means that the probability of being missing is the same for all cases in the data set. This assumption effectively implies that the causes of the missing data

are unrelated to the data.
MAR is based on the assumption that the probability of being missing is the same only within groups defined by the observed data. MAR is usually a much broader group and more realistic than MCAR. If neither the MCAR nor the MAR assumption holds, then we are dealing with NMAR. NMAR means that the probability of missing values depends on the missing values and is not random, so the probability of being missing varies for reasons that are unknown to us.
These are also regarded as nonignorable missingness, and this problem cannot be resolved by direct imputation so it has to be factored into the sample design. When the process generating the missing values is not random, then any imputation procedure that fails to take this condition into consideration will be biased (see Rubin, 1976). If, for example, the probability of reporting a unique item's price is dependent on the actual price, then the process generating the missing values for the item's price is nonignorable, and estimates of the mean and variance of the price will be biased if they are calculated from just the observed data.

Under the assumption of MCAR or MAR, one can use different imputation methods, with varying degree of accuracy depending on the imputation method and the type of missingness. Although most imputation techniques will fare well under the assumption of MCAR, it is unrealistic to use it. The widely-applied assumption of MAR is the most practical, while the NMAR assumption requires a different and more complex approach.

1.3 Study overview and objective

Before the deployment of new estimation software in January, 2015, CPI had imputed missing prices at the item-area level, but after the new system was deployed, the imputation of missing prices was switched to the ELI-index PSU level, a lower and more detailed level. In essence, the mean of the reported price relatives (cell relatives) for similar ELIs within the same adjustment group is multiplied by the previous price of the ELI with the missing price to derive imputed current price. On the other hand, imputation at the item-area level is conducted by multiplying the item stratum-area cell relative of similar items with reported prices by the previous price for the item with the missing price. The item's imputed current price is then used to calculate the item-area index (see the section that follows for details).

One of the main objectives of this study is to identify and justify the reasons for imputing at the ELI-index PSU level as opposed to imputing at the item-area level.
One way to meet this goal is to conduct a simulation study that compares the estimates derived from the imputation of missing prices at the ELI-PSU level versus at the item-area level.

Section 2 of this study explains the procedure used to create the data sets used in the simulation study. Section 3 explains in detail the evaluation method used in the study and the reason it is used. Section 4 analyzes the results of the study. Section 5 is the conclusion and gives our recommendation based on the analysis of the study. Finally, section 6 presents the main references used in the study.

2. Simulation Procedure

For the purpose of simplicity, this study is based on the previous CPI area design that was used until the end of December 2017. Hence, in this study we used about 8000 basic indexes from 211 item strata and 38 geographic areas comprising 87 index PSUs. The data

for this study is from January 2015 through December 2017. We use monthly and bimonthly price quotes from the CPI Research database for the 36-month period from January, 2015 through December, 2017 at the basic item level for the All US area. Imputation at the ELI-index PSU level involves 87 index PSUs, while imputation at the item-area level involves 38 index areas.

2.1 Creating data sets for the imputation

Step 1: Create a "complete" data set and a "mixed" data set for each bimonthly and monthly data set:

- Select all ELI quotes that are active for index calculation in each month of the study, including quotes with imputed prices as a "mixed data set"
- Identify and compute the proportion of quotes with imputed prices at the 80th percentile and the upper 20th percentile for the mixed data set to roughly establish the rate of missingness.
- Select all quotes with observed prices from the mixed data to create a "complete" data set;

Step 2: Using the corresponding "complete" data set, create a pair of six "incomplete" data sets (A and B) for both monthly and bimonthly data as follows:

- For group A, randomly delete the prices for 5%, 10%, and 15% of all the quotes to create the corresponding incomplete data sets with 5, 10 and 15 percent missing prices. We refer to this as the Ordinary random deletion method. Our missingness assumption is MCAR.
- We create data sets for group B by randomly deleting prices after adjusting our simulation values (5%, 10%, and 15%) with the rates of missingness calculated in Step 1 above. Hence, the simulation values are adjusted in proportion to the calculated rate of missingness in our "mixed" data set. We refer to this as "Random deletion proportional to price." See Appendix A for the calculation details. The assumption is that prices could be missing at random (MAR) or not missing at random (NMAR).

Step 3: Impute the missing prices for the incomplete data sets at both the ELI-PSU level and at the item-area level.

The missing data are imputed with cell relative imputation method as follows:

- Calculate the average price relative for each cell using the price relatives computed from the observed prices for similar observations within the cell,
- Multiply the computed average price relative by the previous price for each observation with a missing price to get the imputed current price for the observation.
- Use the imputed price to calculate a price relative for the observation in question.
- The process is repeated for every observation with a missing price and for all the incomplete data sets.

Hence,

$$Cell_j\ relative\ =\ \Pi_{i=1}^{n-1}\left(\frac{p_{ij,t}}{p_{ij,t-1}}\right)^{w}\ ,given\ that\ i=1,2,,n;$$

Where,

$ELI\ quote\ relative, ij\ =\ \frac{p_{ij,t}}{p_{ij,t-1}}$, is the price relative;

W is the Final sample weight;

n is the number of observations i in $Cell\ j$;

t is the current month, and t-1 is the prior month.

Here, to impute the price for observation i in month t with cell relative, and assuming $Cell\ j$ has one observation with missing price $p_{ij,t}^{i*}$ then $p_{ij,t}^{*} = Cell_j\ relative * p_{ij,t-1}$

Imputing at the ELI-PSU level uses the calculated cell relative for observations within the same ELI-PSU group, while imputing at the item-area level uses the calculated item-area relatives for observations within the same item-area group. For the complete data sets, we simply calculate their price relatives.

As a result, we derived the following data sets with calculated price relatives:

- A pair of Complete data – data sets of observations with price relatives of only the observed price quotes (monthly, bimonthly);
- Six pairs of imputed data sets – mix of observed and missing data
    - Incomplete data created by the Ordinary random deletion method
        - 5% deletion (monthly, bimonthly)
        - 10% deletion (monthly, bimonthly)
        - 15% deletion (monthly, bimonthly)
    - Incomplete data created by the Random deletion proportional to price method
        - 5% deletion (monthly, bimonthly)
        - 10% deletion (monthly, bimonthly)
        - 15% deletion (monthly, bimonthly)

3. Evaluation Method

Ideally, we are comparing the imputed value with the value we would have gotten if we had a complete data with all the prices reported. As such, our point of reference is the estimates from the "complete" data set. To evaluate the two methods, we compare their performance in terms of not only how predictive the imputed values are compared to the true values (predictive accuracy), but, most importantly, we compare the two methods fare in terms of the estimation accuracy. Thus, the main focus is to examine the performance and accuracy of imputing at the ELI-PSU level versus at the item-area level. The evaluations are based on the average monthly or bimonthly price relatives over 36 months for ELI-PSU and Item-area group levels.

3.1 Measuring the Strength of Association between Imputed and Complete Data

We compute and compare the correlation between the complete and the imputed data at the ELI-PSU versus at the item-index area level to evaluate their predictive accuracy. We use the Pearson correlation coefficient (r) to measure the strength of the association between estimates from the imputed data and the complete data for each level. The method that yields the higher r value (value nearer to 1) is considered a better choice at imputing the missing data.

Correlation estimates are calculated at each ELI-PSU cell and each item-index area cell, but estimates are averaged over all cells for each PSU or index area, and overall estimates for all US are derived.

It is calculated for ELI-PSU i or item-index area j as

$$r_{\theta_{ijA}^{c}\theta_{ijA}^{c}} = \frac{mean(\sum_{t=1}^{T} \theta_{ijA,t,t-k}^{c}\,\theta_{ijA,t,t-k}^{i}) - \hat{\theta}_{ijA}^{c}\hat{\theta}_{ijA}^{i}}{S(\theta_{ijA}^{c})S(\theta_{ijA}^{i})}$$

Where,

$\hat{\theta}^{c}_{ijA}$, $\hat{\theta}^{i}_{ijA}$ is the mean price relatives over the study period for the complete and the imputed data respectively;

$S(\theta^{c}_{ijA}) = \frac{\sqrt{\sum_{t=1}^{T}\left(\theta^{c}_{ijA,t,t-k} - \hat{\theta}^{c}_{ijA}\right)^{2}}}{T}$ - is the standard deviation for the complete data for ELI – PSU or item-index area over the study period;

$S(\theta^{i}_{ijA}) = \frac{\sqrt{\sum_{t=1}^{T}\left(\theta^{i}_{ijA,t,t-k} - \hat{\theta}^{i}_{ijA}\right)^{2}}}{T}$ - is the naïve standard deviation for the imputed data for ELI – PSU or item-index area over the study period;
T – Number of months involved in the calculation within the study period.

3.2 Measure of Bias and Accuracy

For the purpose of evaluating and comparing the imputation performance between the ELI-PSU level and the item-Area level, we measure the degree of bias and precision for the two levels.

3.2.1 Absolute relative bias

We calculate the relative bias and the absolute relative bias for the two levels. The relative bias is the relative difference in means of the imputed and the complete data for each method. In order words, it is the difference between the averages over the study period of the price relatives of the imputed data and the complete data divided by the average over the study period the price relative of the complete data. The absolute relative bias is derived by taking the absolute value of the relative bias. Both of these estimates express bias relative to the mean of the complete data and serve as a good measure of the degree of biasedness in imputing at the ELI-PSU level versus imputing at the item-area level. The better imputation level is the one that produces a lower relative or absolute relative bias value. The relative bias for an ELI-PSU or item-index area is calculated as follows:

$Relative\ \beta(\Theta)_{iA} = \frac{\beta(\Theta)_{i,jA}}{\hat{\theta}^{c}_{i,jA}} = \frac{\hat{\theta}^{i}_{i,jA} - \hat{\theta}^{c}_{i,jA}}{\hat{\theta}^{c}_{i,jA}}$, and

$Abs.Relative\ \beta(\Theta)_{i,jA} = \left|\frac{\beta(\Theta)_{i,jA}}{\hat{\theta}^{c}_{i,jA}}\right|$

Where,
$\beta(\Theta)_{i,jA} = \hat{\theta}^{i}_{i,jA} - \hat{\theta}^{c}_{i,jA}$ is the bias,
$\hat{\theta}^{c}_{i,jA}$, is the mean price relative over the study period for the complete data for ELI *i in PSU A or item j* in index area A;

$\hat{\theta}^{i}_{i,jA}$, is the mean price relative over the study period for the imputed data for ELI i *or item j* in PSU A.

3.2.2 Normalized root mean square deviation (NRMSD)

To measure the precision of our imputed data at the ELI-PSU level versus at the item-area level, we use normalized the root mean square deviation (NRMSD). It is the standardized root mean square difference between the imputed and the complete data, and is calculated for both imputation levels. The values of NRMSD for both levels are compared, and the one with a smaller value fares better for imputing the CPI missing prices. Normalized root mean square deviation is calculated as

$$NRMSD\left(\theta_{ijA}^{c}\right) = \sqrt{\frac{\left(\hat{\theta}_{ijA}^{i} - \hat{\theta}_{ijA}^{c}\right)^{2}}{Var\left(\theta_{ijA}^{c}\right)}}$$

$Var\left(\theta_{ijA}^{c}\right)$ - Variance of price relatives for complete data (bimonthly or monthly) for ELI i in PSU A or for item j for index area A over the study period.

4. Analysis of the output results

The values of the calculated estimates are tabulated and analyzed. The tabulated results are grouped based on the method used to simulate the missing data: missing data simulated by ordinary random deletion (will be referred to as missing at random), or missing data simulated by random deletion and proportional to price (will be referred to as missing at random and proportional to price). The tables below show the calculated statistical estimates for imputed and complete data at the two levels of imputation (ELI-PSU, item-area).

Table 1: Overall Summary Statistics for ELI-PSU vs Item-Index Area Imputation Level (Using random deletion to simulate missing data.)

FREQUENCY	Cell Level	FULLY OBSERVED VALUES (COMPLETE DATA)			IMPUTED DATA								
					5%			10%			15%		
		N	Mean	SE	R	Mean	SE	R	Mean	SE	R	Mean	SE
BIMONTHLY	ELI-PSU	149557	1.00824	0.02792	0.9092	1.0086	0.028302	0.88037	1.008791	0.028577	0.85284	1.009005	0.02865
	ITEM-AREA	60657	1.00282	0.01472	0.9532	1.0029	0.014937	0.92711	1.003018	0.015213	0.90306	1.003252	0.01548
	DIFF (GAP)		***0.00543***	***0.0132***	***-0.044***	***0.0057***	***0.01336***	***-0.0467***	***0.00577***	***0.013364***	***-0.0502***	***0.005753***	***0.01317***
MONTHLY	ELI-PSU	276405	1.00264	0.01101	0.9528	1.00276	0.0112	0.92731	1.002882	0.011421	0.89944	1.003027	0.01163
	ITEM-AREA	117872	1.00123	0.00791	0.9719	1.00129	0.008073	0.94918	1.00135	0.008275	0.92305	1.001494	0.00847
	DIFF (GAP)		***0.00141***	***0.0031***	***-0.019***	***0.00146***	***0.00313***	***-0.0219***	***0.00153***	***0.003146***	***-0.0236***	***0.001533***	***0.00316***

Table 1 shows that the imputed data are highly correlated with the complete data for both ELI-PSU imputation and the item-index area imputation. The correlations for imputed monthly data are higher than for imputed bimonthly data. The table shows that the estimated values of the mean and the standard errors for the imputed data for both methods are very close to the calculated "true values" for the complete data. The data imputed at the Item-Area level are slightly more correlated to the corresponding complete data than the ones imputed at the ELI-PSU level. Similarly, the values of calculated estimates of the mean and standard errors are closer to their corresponding true values computed for the complete data at the item-index area level than they are at the ELI-PSU level. The table also shows that as the amount of missing data increases, so also does the differences between the values of these estimates for the imputed data and the calculated values for the complete data.

Table 2: Overall Summary Statistics for ELI-PSU vs Item-Index Area Imputation (Using data missing at random and proportional to price.)

FREQUENCY	Cell Level	FULLY OBSERVED VALUES (COMPLETE DATA)			IMPUTED DATA								
					5%			10%			15%		
		N	Mean	SE	R	Mean	SE	R	Mean	SE	R	Mean	SE
BIMONTHLY	ELI-PSU	149557	1.00824	0.02792	0.9357	1.00829	0.027959	0.93372	1.008362	0.028013	0.9318	1.008382	0.02806
	ITEM-AREA	60657	1.00282	0.01472	0.975	1.00279	0.014787	0.97022	1.002841	0.014852	0.96615	1.002801	0.01487
	DIFF (GAP)		***0.00543***	***0.0132***	***-0.039***	***0.00549***	***0.01317***	***-0.0365***	***0.00552***	***0.013161***	***-0.0344***	***0.005581***	***0.0132***
MONTHLY	ELI-PSU	276405	1.00264	0.01101	0.9764	1.00265	0.011025	0.97457	1.002653	0.011042	0.97291	1.002672	0.01106
	ITEM-AREA	117872	1.00123	0.00791	0.9891	1.00126	0.007957	0.98472	1.001256	0.007992	0.98069	1.001259	0.00803
	DIFF (GAP)		***0.00141***	***0.0031***	***-0.013***	***0.00139***	***0.00307***	***-0.0101***	***0.0014***	***0.00305***	***-0.0078***	***0.001413***	***0.00303***

The results in Table 2 are based on the data from simulating missing pricing at random and in proportion to price in the research database. The results are consistent with those in Table 1. The correlation coefficients for both monthly and bimonthly are very high for all the imputed data for the two imputation levels. In contrast to the results in Table 1, the calculated statistics are not as sensitive to the increase in the missingness in price.

Tables 3a and 3b show estimates of the overall relative bias between the imputed data sets and their corresponding complete data. The two results show a divergence of performance by the two levels of imputation based on the type of data sets being imputed. For data missing at random, imputed data at the Item-Area level have a lower relative bias and absolute relative bias than those imputed at the ELI-PSU levels whether they are monthly or bimonthly data. But the result is split when using data sets with values missing in proportion to price. Imputations done at the ELI-PSU level have lower relative biases and absolute biases than the ones done at the Item-Area level for the monthly data. For the bimonthly data, imputing at the Item-Area level gives lower values of relative biases than doing so at the ELI-PSU level.

TABLE 3A: Overall Estimates of Relative Bias of Mean and Normalized Root Mean Square Deviation (Using data missing at random.)

Frequency	Imputation	DATA SIZE		RELATIVE BIAS		ABS RELATIVE BIAS		NRMSD	
	Cell Level	ELI-PSU	ITEM-AREA	ELI-PSU	ITEM-AREA	ELI-PSU	ITEM-AREA	ELI-PSU	ITEM-AREA
Bimonthly	5%	149557	60657	-0.0003	-0.0001	0.0033	0.002	0.1467	0.1673
	10%	149557	60657	-0.0006	-0.0002	0.0058	0.0031	0.2559	0.2772
	15%	149557	60657	-0.0008	-0.0004	0.0076	0.004	0.3401	0.36062
	average	***149557***	***60657***	***-0.0006***	***-0.0002***	***0.0056***	***0.003***	***0.2473***	***0.2684***
MONTHLY	5%	276405	117872	-0.0001	-0.0001	0.0017	0.0013	0.1671	0.1774
	10%	276405	117872	-0.0002	-0.0001	0.0027	0.0019	0.2706	0.2761
	15%	276405	117872	-0.0004	-0.0003	0.0035	0.0025	0.3594	0.3642
	Average	***276405***	***117872***	***-0.0002***	***-0.0002***	***0.0026***	***0.0019***	***0.2657***	***0.2726***

Analyzing the results for NRMSD in table 3a and 3b shows that imputing at the ELI-PSU level fares better than imputing at the item-index area level. Although the values of NRMSD are low for all the imputed data, NRMSD values are lower for the data imputed at the ELI-PSU level, and are less than half as high for the simulated data missing at random proportional to price for both monthly and bimonthly data as compared to the same data but imputed at item-index area level. This essentially indicates that imputation carried out at the ELI-PSU level for these data is more accurate than at the item-index area level. The

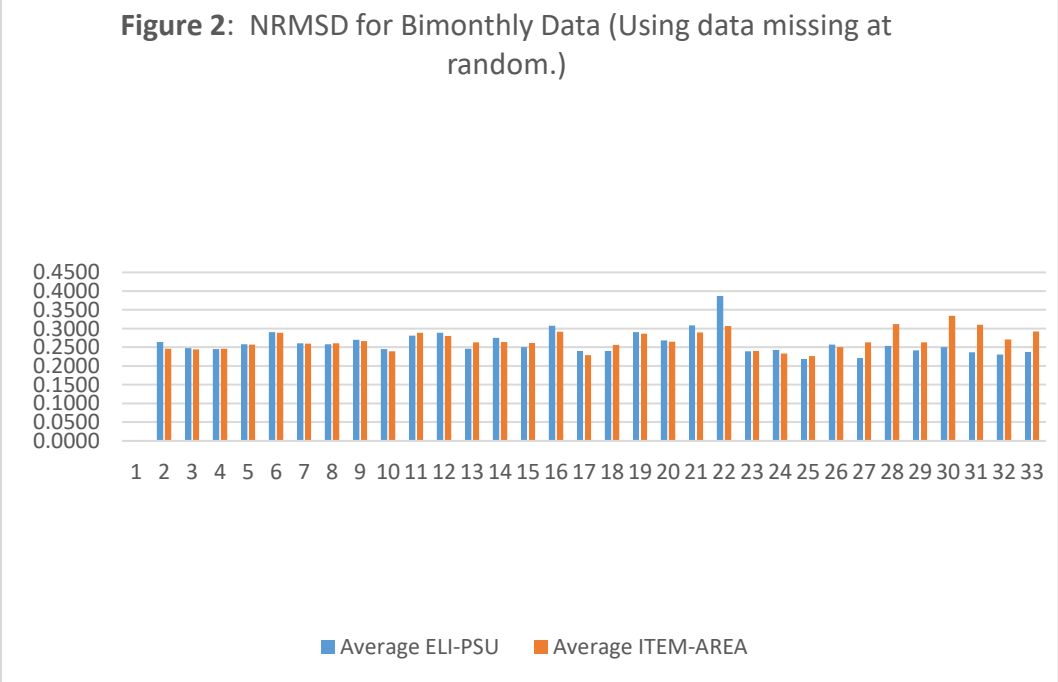
results of a similar analysis conducted for individual PSU and index area mostly mirror the overall results.

TABLE 3B: Overall Estimates of Relative Bias of mean and normalized root square Deviation (data missing at random proportional to price)

Frequency	Imputation Cell Level	Data Size		Relative Bias		Abs. Relative Bias		NRMSD	
		ELI-PSU	ITEM-AREA	ELI-PSU	ITEM-AREA	ELI-PSU	ITEM-AREA	ELI-PSU	ITEM-AREA
Bimonthly	5%	149557	60657	-0.0001	0.0000	0.0006	0.0008	0.0254	0.0621
	10%	149557	60657	-0.0001	0.0000	0.0009	0.0013	0.0429	0.1003
	15%	149557	60657	-0.0001	0.0000	0.0012	0.0015	0.057	0.1264
	average	***149557***	***60657***	***-0.0001***	***0.0000***	***0.0009***	***0.0012***	***0.0418***	***0.096***
Monthly	5%	276405	117872	0.0000	0.0000	0.0003	0.0005	0.0308	0.07
	10%	276405	117872	0.0000	0.0000	0.0005	0.0007	0.048	0.1051
	15%	276405	117872	0.0000	0.0000	0.0006	0.0009	0.0604	0.1303
	Average	***276405***	***1E+05***	***0.0000***	***0.0000***	***0.0004***	***0.0007***	***0.0464***	***0.102***

The tables show the output results for normalized root mean square deviations by index area for the two levels of imputation process: ELI-PSU level imputation and Item-Area level imputation. For the purpose of consistency, we aggregated the ELI-PSU level results for non-self-representing index PSUs to index Area level to match with the Item-Area level.

The results compare the efficiency of the two imputation levels by accessing their normalized root mean square deviation between the data with imputed values and the data with fully observed values. Both results for monthly and bimonthly data show that the two levels performed fairly well, but imputation at the ELI-PSU level outperformed the item-Area level for all the simulated data.



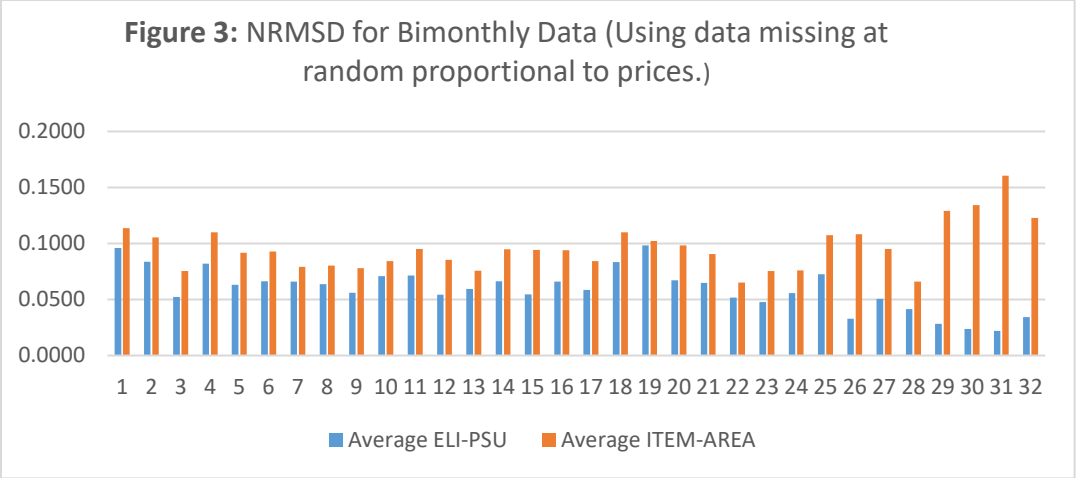
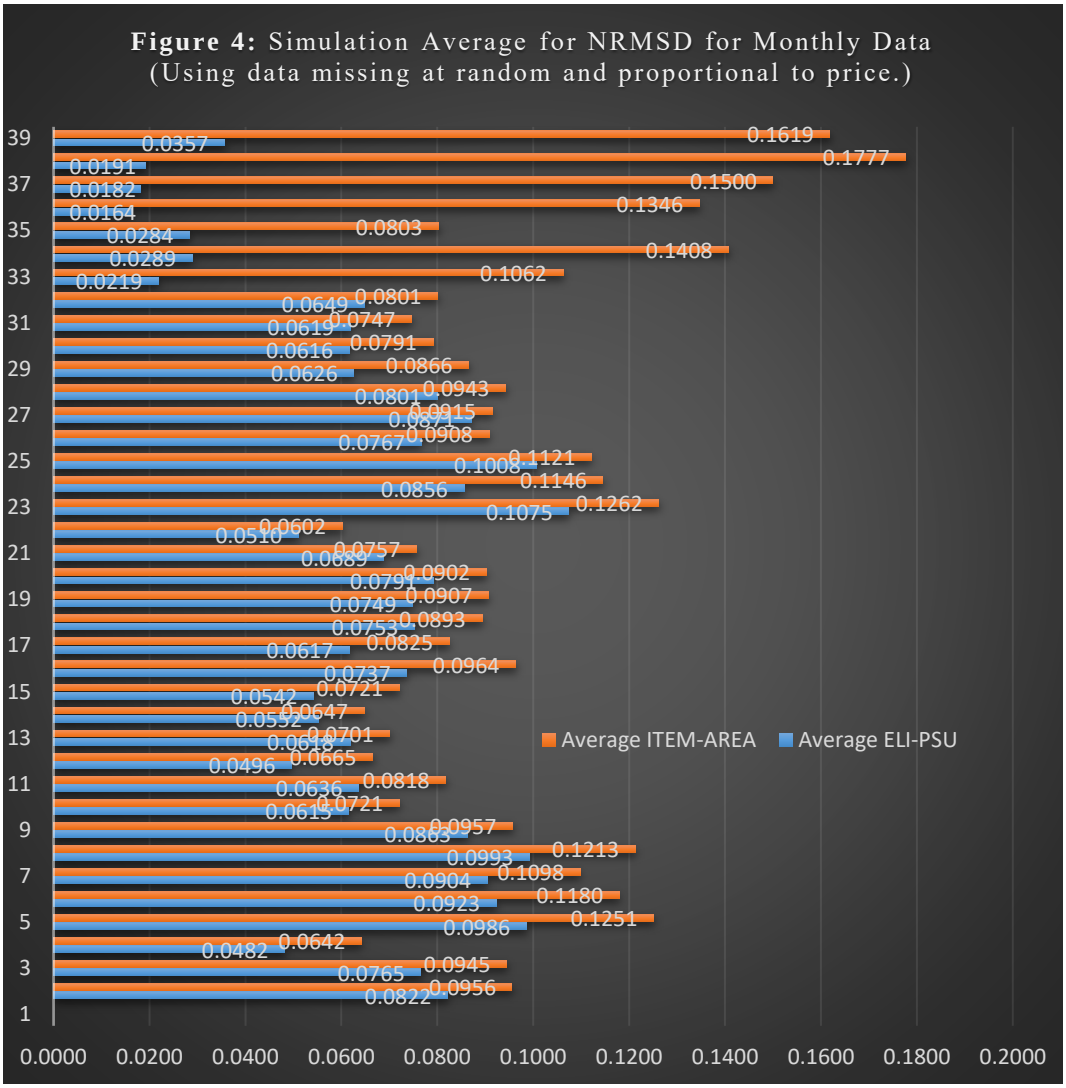
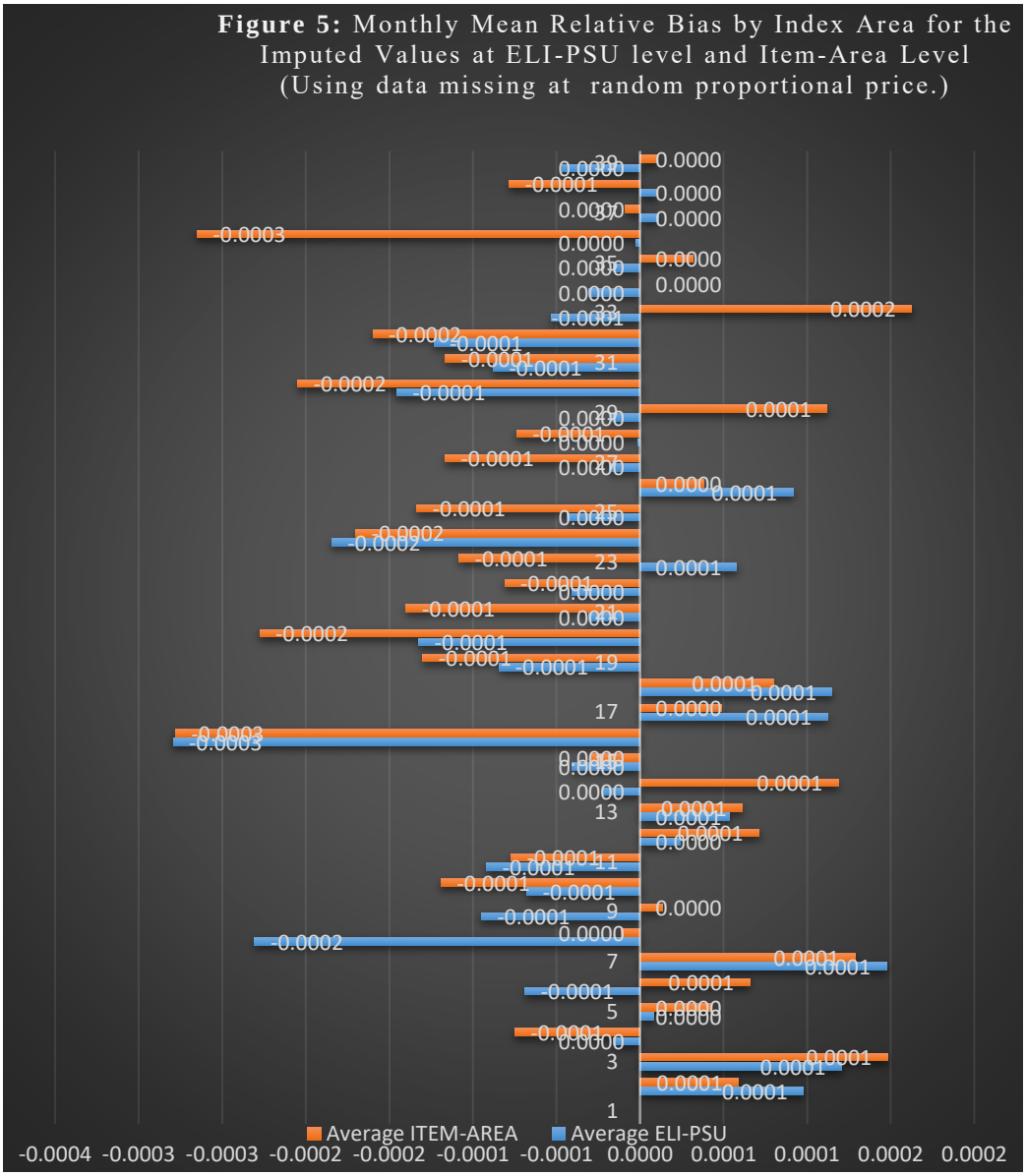


Figure 2 through 5 are graphical representation of some of the results for the PSUs and the index areas.





5. Conclusion and recommendation

The overall results of the study show that the current CPI process of imputing at the ELI-PSU level performed better than the previous method of imputing at the Item-Area level.

For the "ordinary random deletion method," in which we assumed that missing prices are MCAR, the summary statistics of the imputed data show no discernible distinction between the two levels of imputation, and, as expected, the imputed data at both levels are highly correlated with the corresponding complete data.

The results for the relative bias show a very mixed picture, with a slight advantage to the Item-Area level imputation for data missing at random (MCAR assumption), although the difference is minuscule and could easily be attributed to noise created by the rounding

errors during the aggregation process from ELI-PSU to item-index area. However, this result also shows that imputing at the ELI-PSU level outperformed imputing at the Item-Area level for prices missing at random and proportional to price, where our assumption is based on MAR and the possibility of NMAR.

Imputation precision or performance evaluation of the two levels of imputation show that the ELI-PSU imputation level outperformed the Item-index area imputation level at all levels of missingness. Estimates of the normalized root mean square deviation (NRMSD) of ELI-PSU level imputation are smaller than that of Item-Area level imputation. Also, we observed that this result is similar for both types of simulated incomplete data (by ordinary randomly deleted data or randomly deleted proportional to prices). This is a clear advantage for implementing ELI-PSU level imputation rather than Item-Area imputation.

However, it is important to mention that in most of the other instances, the difference in results between imputing at the two levels is not large.

The overall results show that there is no meaningful distinction between our two assumptions of missingness: data missing at random and missing at random proportional to price. The study did not show any clear evidence of a difference between the two assumptions of missingness.

Although the results of the study show that both imputation procedures performed reasonably well, it would not be a mistake if we chose to use the ELI-PSU level imputation method.

What next?

- This study is done based on simulating missingness at three different percentages of missing values - 5, 10, and 15 percent. The scope of the study could be expanded to include simulating missing values at 20, 30, or higher percentages to see what would happen.
- An experimental study could be done to examine the impact of using different imputation methods such as hot deck, multiple imputation on the CPI.
- Other questions to be considered are:
    - What should be done to improve bimonthly data that are carried forward during off-cycles, which is an implicit form of imputation?
    - How could we alleviate some of the obvious drawbacks presented by the use of cell mean imputation because of its tendency to diminish estimates of means, variance and standard errors?

References

"Alternative Paradigms for the Analysis of Imputed Survey Data" by Robert E. Fay, Journal of the American Statistical Association, Vol. 91, No. 434 (Jun. 1996).

"Applying Mass Imputation Using the Schools and Staffing Survey Data" by Steven Kaufman et al.

"Applied Missing Data Analysis" by Craig Enders (2010).

"Assessing the impact of missing data on hospital performance profiling" by michael p. Thompson (2015).

BLS Handbooks of Methods.

“Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments” by Magalie Celton et al. BMC Genomics 2010.

“Estimating Missing Prices in the Producer Price Index” by Onimissi M. Sheidu, BLS (2005).

“Flexible Imputation of Missing Data” by Stef van Buuren (2012).

“Fully Conditional Specification” by Stef van Buuren, page 267 – 258, Handbook of Missing Data Method.

“Imputation of Missing Values, Seasonal Products, and Quality Change” by Jessica Penvose, BLS (2016).

“Longitudinal Imputation of SIPP Food Stamp Benefits No. 9406” by Antoinette Trembay, U.S. Bureau of Census.

“Multiple imputation for Missing Data: A Cautionary Tale” by Paul D. Allison, University of Pennsylvania.

“Multiple imputation for nonresponse in surveys” by Rubin (1987). New York, N.Y: John Wiley and Sons.

“Multiple imputation of missing income data in the national health interview survey” by Raghunathan T. and Schenker, N. Journal of the American Statistical Association, 101, 924-933.

“On Variance Estimation with Imputed Survey Data” by I. N.K. Rao, Journal of the American Statistical Association, Vol. 91, No. 434 (Jun. 1996).

“Some General Guidelines for Choosing Missing Data Handling Methods” in Educational Reseach by Jehanzeb R. Cheema. University of Illinois at Urbana Champaign (2014).

“Some Remarks on the data imputation using “Missforest” method” by Malgorzata Misztal, Department of Statistical Method, University of Lodz (2013).

“Systematic Assessment of Imputation Performance using 1000 Genomes Reference Panels” by Qian Liv et al., Oxford Journals: Briefing Bioinformatics (2015).