ON COLD-DECK IMPUTATION WITH DATA QUALITY IMPROVEMENT USING SIMULATION MODEL

SEAN WANG
AMERICAN ENTERPRISE INSTITUTE FOR PUBLIC POLICY RESEARCH

ABSTRACT. This article introduces a way of imputing missing values occurred in the public use tax file. The missing data mechanism can be described as missing not at random, and, based on such mechanism, we use some external data source, the consumer expenditure survey, to perform Cold-Deck imputation. Before units in the two files are being matched, tax-calculator, a tax simulation model developed by Open Source Policy Center, is used to improve data quality from the survey by discarding units in the survey data that are considered non-imputation recipients in the original file. Nearest neighbors algorithm is used to perform the imputation, while cross validation takes care of within-model parameter selection. Occurrence frequency measure is also introduced to quantify the Euclidean distance for non-ordinal categorial variable. Finally, we check robustness to ensure our imputation work will not break current missing mechanism, and discuss imputed results.

1 Introduction

This article intends to introduce a way of imputing missing values for 6 itemized deduction variables within 2009 public use tax file (09 puf) provided by Internal Revenue System (IRS), where the pattern of missing data can be described as missing not at random (MNAR). More specifically, the reason for missing depends on the unseen observations themselves, even when we account for all the available information observed. In our case, these variables are lacking those observations (non-itemizers) because, roughly speaking, the summation of these missing variables is less than a certain threshold, the standard deduction amount, that leads to missingness.

There are various approaches to impute missing data of such pattern. As discussed in the publications from The Joint Committee on Taxation [1] and Bureau of Labor Statistics [2], these approaches suggest matching or merging with external survey data files, like the Consumer Expenditure Survey (CEX). We take such Cold-Deck imputation method that uses the latest release of CEX data, where, thanks to a micro-simulation model, tax-calculator, developed by open source policy center, the dataset is being "cleaned" in the way that it is unlikely to contain non-itemizers, and thus improves the quality of our donor. In order

Date: April 14, 2017.

to do so, consumer units (CU) in CEX are divided into one or two filing units based on their number of earners and marital status. If two filers were ever being split from one CU, then their expenditure, number of dependents and other variables will also be split according to their respect earnings. Married filers in CEX are being assigned to either joint filers or separate filers by means of stratified sampling, where proportionate allocation is applied to reflect the "joint vs separate" ratio in the dataset to be matched.

Nearest neighbor algorithm is implemented to perform imputation, where we introduced the occurrence frequency measure to obtain the Euclidean distance for non-ordinal categorial variable, and thus to calculate the between-record distances. Moreover, cross validation is being used to access within-algorithm model goodness.

To ensure that the missing mechanism is not violated after imputation, robustness is checked for each imputed record and we take treatment on those "wild" ones.

2 Data

I. Missing Mechanism in Original Dataset

In the 09 puf, missingness follows a pattern called MNAR, where the probability of missing values depends on the variables that are missing themselves. That is, let Y be the matrix representation of the data, M be the indicator matrix of missing data, and θ be the unknown parameters, then

$$\mathbb{P}(M \mid Y, \theta) = \mathbb{P}(M \mid Y_{missing}, \theta).$$

For MNAR data, selection models or pattern-mixture models can be used to impute the data. Methods like these, however, require the distribution of missingness to be explicitly specified. In our case, although the mechanism of missingness, as well as the threshold that is equipped with each filing unit, are both known, the restraints are not enough for us to assume a specific underlying distributions for each of those various missing variables. In light of this, we take the Cold-Deck imputation approach that uses external data sources. Analogous to Hot-Deck imputation, missing values for a non-respondent (recipient) are replaced with observed values from a respondent (donor) that is similar to the nonrespondent with respect to characteristics observed in both cases.

II. Structure of External Data Source

The CEX is used as our external donor to impute missing parameters. CEX is helpful because it is the only Federal survey to provide information on a complete range of consumers' expenditures and incomes, as well as the characteristics. It is widely used by economic policymakers examining the impact of policy changes on economic groups, by the Census Bureau as the source of thresholds for the Supplemental Poverty Measure, by businesses and academic researchers studying consumers' spending habits and trends, by other Federal agencies, and, perhaps most importantly, to regularly revise the Consumer Price Index market basket of goods and services and their relative importance.

The Consumer Expenditure Survey (CEX) program consists of two surveys, the Quarterly Interview Survey and the Diary Survey, that provide information on buying habits of American consumers. These includes data on their expenditures, incomes, and consumer unit (CU) characteristics for both families and single consumers. In order to perform the data quality improvement and the imputation, we break down CU data into member level (filing units), and transfer quarter information (Diary Survey is not used) into annual data.

III. Data Manipulation

◦ Data Cleaning

Not all records in the CEX are being used. In particular:

- CUs with more than one reference person, meaning units contain more than one family, are being excluded.
- Surviving spouse units, that is CUs where the reference person is widowed or married but with no spouse entry, are being excluded.
- CUs with zero total earnings are being excluded.

Although the data to be imputed, 09 puf, does contain widowed records, such status is censored in a way that no further information could be obtained. To avoid introducing extra noise, we simplify our assumptions by discarding records like this.

- Temporal Transformation

  As mentioned earlier, quarterly data provided by CEX need to be transferred into annual data, since this is the format used in the 09 puf. We first check whether the variables we are using exhibit seasonality by looking at their respective averaged values across different quarters.

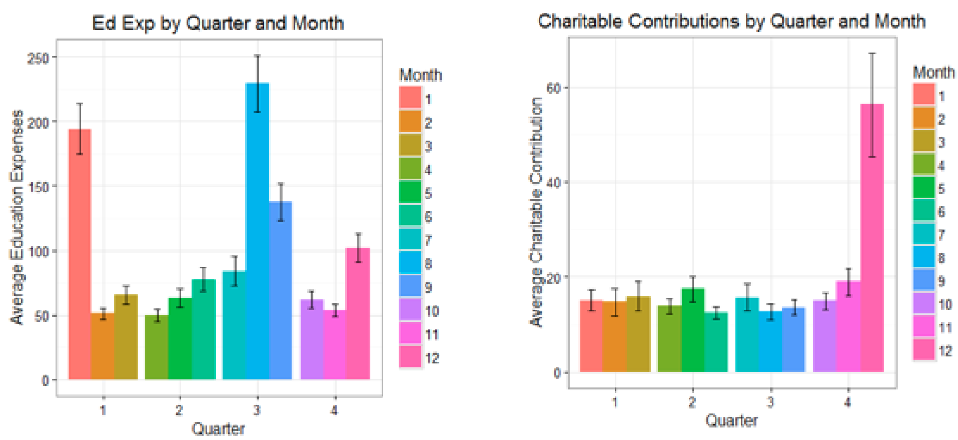

FIGURE 1. Before seasonality treatment

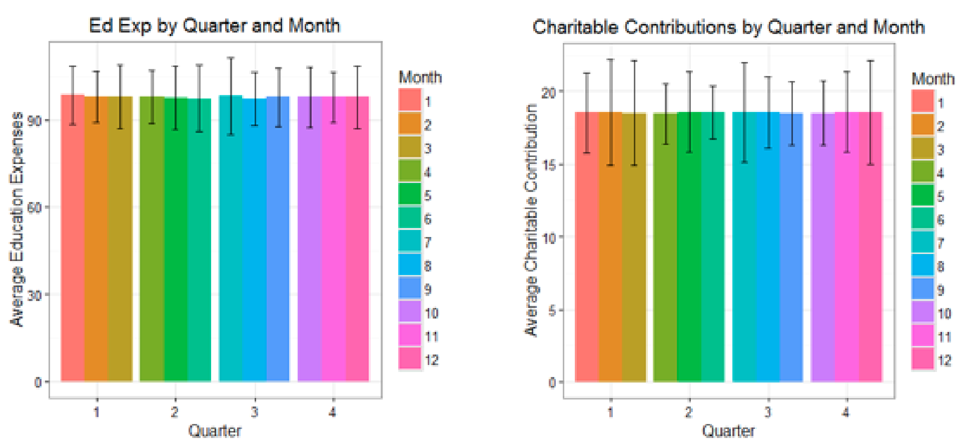

FIGURE 2. After seasonality treatment

In regards to expenditure variables extracted from MTAB files, for which monthly data is available, we face the same problem due to the fact that most of this monthly data is only available for one quarter. After plotting average monthly values for the concerned variables, we find Education Expenditure (Tuition and fees) and Charitable Contributions to exhibit noticeable seasonality, as indicated in Figure 1. Due to our method of estimating annual data from these variables, seasonality poses a risk of either overestimation or underestimation.

We decided to address this seasonality issue by giving each months specific weights according to their mean distribution, and recode the data to reflect each months respective weight. To acquire these weights, the mean of monthly averages are computed (for now, lets call this the mean of means). We then compute each months weight by dividing the mean of means by that months average, so that a higher-than-average month will be assigned a smaller weight. Figure 2 displays the distribution of monthly averages after this treatment, which got rid of any seasonality previously observed in the variables.

- Annualizing Data

  We proceed to estimate the annual expense by employing selective data modification methods. Although quarterly recorded, some variables in the Consumer Expenditure Survey, such as salary/wage or pensions, actually contain annual data. For these variables, we only keep observations from the latest interview month since the data refers to earnings acquired 12 months prior to the time of interview. For quarter data, we try our best to avoid seasonality issues by using imputations from every available quarter, instead of estimating annual data from quadrupling one quarters observations.

  Let q be the number of quarters where the CU/members data is available, d_i be the quarter data recorded at quarter i, and D be the annual data. We have:

$$D = \sum_i 4 * \frac{d_i}{q}.$$

  Variables constructed from MTAB files are available in the form of monthly data, and annualized by summing available monthly data into quarter data, then multiplying the result by 4. And this yields the desired annual data.

- Expenditure Allocation

While breaking down CUs into filing units, we also allocate the expenditure of CUs to members within each CU, based on their respective earnings. More explicitly, let $\mathcal{T}$ be the total earning of each CU, n be the number of member(s), t be the earning of respective member(s) in that CU, and $E_1, E_2, \ldots, E_k$ be various expenditures. Then

$$E_i^j = \frac{t_i}{\mathcal{T}} * E_j, \text{ where } i \in n,\ j \in k \text{ and } \sum_i t_i = \mathcal{T}.$$

Following such procedures, we are able to obtain member level expenditures.

- Dependency Test

  After obtaining the necessary data for Marital Filing Status among households and family members, we are able to perform the Dependency Test by partly employing Lorez Kueng's methodology from the Cex-TAXSIM [1] project. Dependency is determined via a mix of relationship, age, and self-support tests. In this particular example, we use the threshold of $3,650$ dollars to determine if the members total yearly income is efficient for self-support, both for qualified children and relatives dependency tests. We later impute the number of dependents by looking at each family and their marital filing status. For joint filers, the reference person claims all qualified dependents including the spouse, while separate filers divide the number of dependents by their respective earning capacity.

- Re-sampling Married CUs

  Since the Consumer Expenditure Survey provides no information on specific Marital Filing Status, we resort to stratify sampling to fill in the missing data and compute a new variable named MARS. The categorical variable corresponds to the MARS variable in 09 puf, which documents filing status by integer values ranging from 1 to 4. Single filers (type 1) are identified as consumer units with family size of 1, while Head of households (type 4) are assigned to non-married reference people from CUs with more than one family member but where only one of which makes an income. Divorced households are also listed under type 4.

  Among married households, we impute type 2 (Married filing jointly) and type 3 (Married filing separately) statuses by stratify sampling the CU pool using available 09 puf

[1] More details here: http://users.nber.org/ taxsim/to-taxsim/cex-kueng/

data. First, a ratio of the number of type 2 to type 3 filers is computed for each of four different income brackets in the puf dataset. For each earnings bracket in CEX, a sample of married reference people (representing the household) is selected and assigned Married Filing Jointly (type 2) status, while the rest are treated as separate (type 3) filers. The sample sizes are computed to reflect the ratio of type 2 to type 3 filers in the corresponding income bracket in the puf dataset. To ensure consistency in member and CU level data, sampling and assignment is executed on reference person entries, then assigned to the corresponding spouse.

IV. Data Quality Improvement

After having compatible data, and before moving toward imputing the original dataset, we use a Microsimulation model [2] developed by the Open Source Policy Center to improve the data quality of CEX. The model plays a role here helping us filter the information extracted, where obviously non-similar units, who are unlikely to be non-itemizers, are being discarded from the pool.

The procedure follows that:

(1) Feeding entire CEX data into Tax-Calculator.

(2) Calculating taxes for each record.

(3) Discarding records with high itemized-deduction amount.

We end up with 5689 records, where about 10% records have been dropped from the original CEX dataset.

3 Metric, Algorithm and Weight

I. Occurrence Frequency Metric

In order to measure the distance for categorial variable, we introduce the Occurrence Frequency (OF) measure [3] that assigns mismatches on less frequent values with lower similarity (i.e. larger distance) and mismatches on more frequent values with higher similarity (i.e. smaller distance). Our reason for selecting the OF measure is because,

[2]Full model can be found here: https://github.com/open-source-economics/Tax-Calculator

[3]The measure is discussed in [3] under similarity setting, while we use the distance form under similarity-distance transformation $\mathrm{sim} = \frac{1}{1+\mathrm{dist}}$.

as discussed in [3], the OF measure yields robust results against outliers in terms of outlier detection performance.

Distance for OF measure is defined

$$d(X,Y) = \begin{cases} 0 & \text{if } X = Y \\ \log(\frac{N_X}{f_X(X)}) * \log(\frac{N_Y}{f_Y(Y)}) & \text{o.w.} \end{cases},$$

where X and Y are our objects, N_X is total number of records in the dataset where X belongs, and let $f_X(X)$ be the occurrence frequency of outcome X in the dataset where X belongs. Similarly, we define N_Y and $f_Y(Y)$.

II. *K*-Nearest Neighbor

Nearest neighbor [4] expects the conditional probabilities to be almost locally constant. We use K nearest records in terms of standard Euclidean metric to measure similarity and perform the imputation. Figure 3 illustrates how different K might affect our decision.

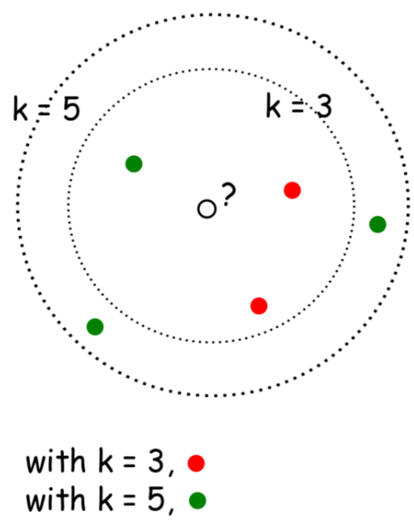

FIGURE 3. Nearest Neighbor with Different Options of K

In our case, we use wage/income, number of dependent(s), and transformed metric for respective marital status, the occurrence frequency measure, as our input variables. Based on distance determined by the three variables under standard Euclidean setting with proper weight for each component (details in Weighting section), we are able to

[4]More details about this methodology can be found in Chapter 2 of [4].

select K nearest points in the CEX data set when some record (to be imputed) is given. Once K points have been selected, for each missing expenditure of each given record, we obtain an averaged amount based on these K points and impute this missing variable. For one given record, these K unique points can be used to complete the imputation, by repeating the previous step for respective missing variables.

III. Weighting Strategy

Since variables that we used to calculate the distance have different scales, and in order to ensure that each variable effects the distance in a reasonable way, we assign different weights to different variables. Weights are determined in the following way:

$$\begin{cases} \omega_1 + \omega_2 + \omega_3 & = 1 \\ \omega_1 : \omega_2 : \omega_3 & = \frac{1}{\overline{\mu_1}} : \frac{1}{\overline{\mu_2}} : \frac{1}{\overline{\mu_3}} \end{cases},$$

where ω_1, ω_2 and ω_3 are weights for wage, total number of exemptions and transformed martial status respectively. And we define $\overline{\mu_1}$ to be

$$\overline{\mu_1} = \frac{1}{|N_x|} \frac{1}{|N_y|} \sum_{x \in N_x} \sum_{y \in N_y} (\text{Wage}_x - \text{Wage}_y)^2,$$

the averaged square difference of all possible enumerations between donor dataset N_x and recipient dataset N_y for wage variable. Similarly, we define $\overline{\mu_2}$ and $\overline{\mu_3}$. Adopting weights in this way will allow us to obtain reasonable distances in the way that no single variable would overwhelm, while maintaining the original information in the datasets as much as possible.

4 Model Selection

We use cross validation [5] to perform within-model selection for nearest neighbor algorithm, and to decide what's the "best" K to use. Also, a test set has been separated from CEX (our donor population) to perform evaluations for variations of the algorithm. Note that test set will not be used anywhere else except final model assessment to ensure the "sanity" of our assessment.

[5] Information about cross validation can be found in Chapter 7 of [4].

I. N-Fold Cross Validation

We resort to cross validation to determine what is the “best” K to use. Ideally, we set aside a validation set (within the training population that does not include test set) to assess the performance of our prediction model. Given scarce data, N-fold cross validation is used to finesse this problem, where part of the available training donors are being used to fit the model, and a different part to test it.

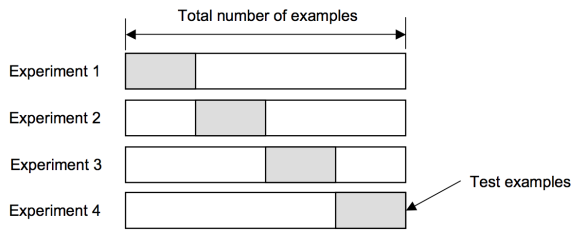

FIGURE 4. Example of 4-Fold Cross Validation

Figure 4 gives an example of 4-fold cross validation. Note that:

- N-fold cross validation requires N experiments.
- After each “shuffle”, validation set of each experiment remains the same.
- Model assessment is based on combined prediction error from all experiments.

We now proceed to assess goodness of different models, that is, different options of K, using 5-fold cross validation.

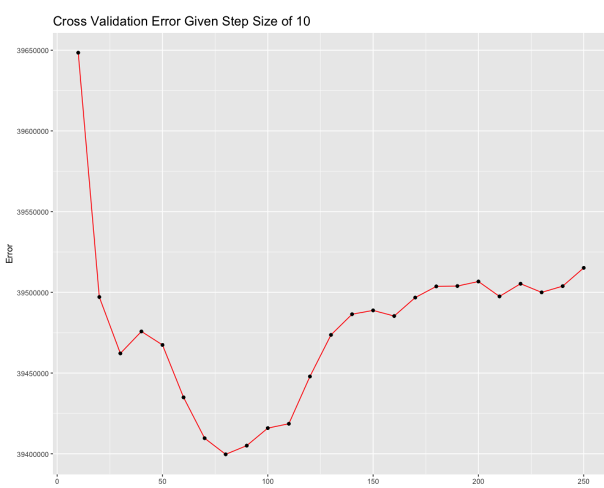

FIGURE 5. Large Step Size for K from 0 to 250

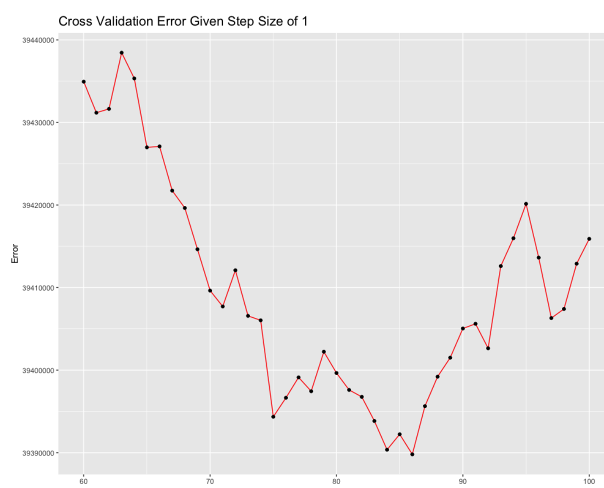

FIGURE 6. Small Step Size for K from 60 to 100

Figure 5 shows sum of squared error against different options of K, where each step is of size 10. The bowl-shaped curve is known as the bias-variance trade-off, where bias dominates over variance when K is small, and vice versa when K is large. Thus, as suggested by the plot, appropriate choices of K would yield lower error. Figure 6 indicates that cross validation method is locally unstable when given small step size, and thus only suggests an ambiguous choice of K. This, however, should not be a concern since slightly different choices of K, say $K = 82$ and $K = 86$, would result in rather insignificant validation difference. We hence pick $K = 80$ for our algorithm.

II. Test Error and Comparison

In order to show that occurrence frequency measure and our weighting strategy do improve the matching accuracy, we go ahead and compare four different models:

- The naive model that includes input variables under normalization with uniformed weight for each of the variable.
- The OF model that introduces OF measure for the categorial variable, but normalized input variables are fitted with uniformed weight for each of the variable.
- The weighted model that evaluates proper weight for each variable and intakes raw (that is non-normalized) input variable information.
- The full model that uses OF measure as well as weighting evaluation.

Each of the above model undergoes a model-selection process that yields the "best" number of neighbors K to use, which we have already introduced in the previous section. We should expect that the full model outperforms, overall, all other models, and the naive model under-performs. Given this, we use the mean square error from the naive model as baseline, and evaluate the percentage rates each model outperforms the naive model[6].

Figure 7 gives a comprehensive evaluation of model performance on each variable imputed. From the figure, we observe that:

- The occurrence frequency model improved, moderately, the performance in three of the imputed variables.

[6]The percentage rate is given by $\frac{\text{MSE}_{\text{naive}} - \text{MSE}_{\text{other}}}{\text{MSE}_{\text{naive}}}$. When negative, it means the naive model outperforms.

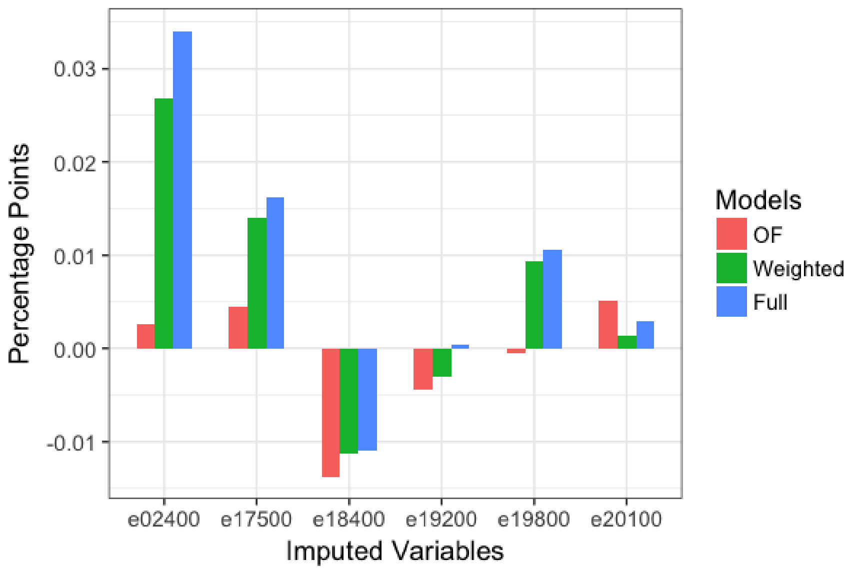

FIGURE 7. Model Performance by Percentage Points

- Full model and Weighted model significantly improved matching performance in three of the imputed variables.
- The full model always outperforms the weighted model.
- For one particular variable, e18400, all three models are outperformed by the naive model over 0.01. We will have more discussions regarding this in later section.

5 Imputed Results

Within the 09 puf dataset, 123, 114 records are considered as recipients. Without taking robustness into consideration, we use the full model and carry out the matching and imputing process. After imputation, the initial imputed puf dataset ends up with 3740 non-valid records that do not pass the robust test. Given this is a rather insignificant portion among all recipients, we simply undo the imputation for those records, instead of using any further treatments. This completes the imputation work.

The robustness test is designed in the way that, after the imputation, the imputed dataset will not break the original missing mechanism record-wise. That is, all non-itemizers (recipients) will remain non-itemizers.

Figure 8 shows the density distribution of each variable imputed for our recipients. The distribution for variable e20100 seems odd in the way that we do not see much variation in the imputed result. Possible causes are included in the discussion section.

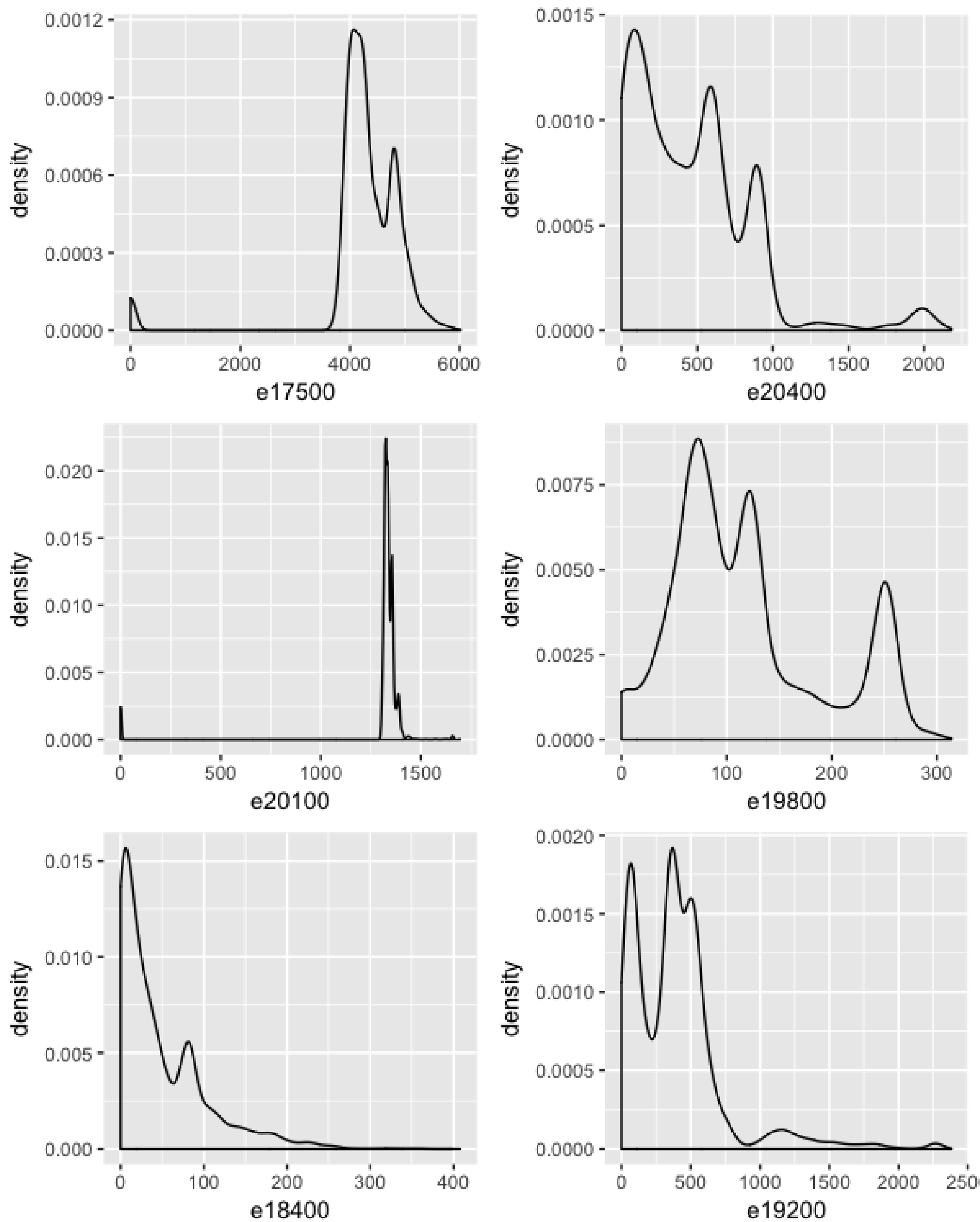

Figure 8. Density Distribution for Imputed Variables

Note that, since we are using 14 CEX, the most recent CEX release, as our donors, while our recipients are in year 2009, those imputed expenses also require a reversed extrapolation process that brings 2014 data into 2009 level.

6 Discussions

Cold-deck imputation requires donor from external source, which is sometimes not easy to obtain. Fortunately, consumer expenditure survey (CEX), as an ideal candidate, offers adequate information on matching similar units, and imputing desired expenses as well. There are, however, a few drawbacks using data source like CEX, and using methodologies like cold-deck imputation.

Given the missing mechanism, cold-deck imputation usually yields outstanding outcome when comparing to some other methodologies. Such method, however, requires maintenance from time to time. When it comes to updates on either donor's side or recipient's side, we would have to update the imputation procedure in order to obtain compatible and sensible results. This can sometimes be challenging, because slight data structural changes, like adding or removing certain variables, would lead to considerable modifications in our current work.

On the other hand, the donor we used itself, CEX, carries a few limitations. An obvious drawback is the sparsity of observations. Using only $5,689$ records to impute $123,114$ records might result in potential bias. In figure 8, we observed that one of our imputed variable, e20100, do not have much variation. A plausible explanation would be the original variable in the donor has low volatility, which turns out to be our case. A straightforward treatment to deal with this issue is bootstrap re-sampling. Pooling previous year CEX releases together will definitely alleviate such bias, but require more work. These possible solutions shall be incorporated in the future as one of major improvements. Another minor drawback is that, since CEX mostly offers information in consumer unit (family) level, potential errors might be introduced while we are interpreting these information into tax filer (individual/couple) level. Without adopting any complicated assumptions, our interim estimation mainly addresses this problem and provides us with compatible data as donor to complete the imputation work.

In terms of modeling, we introduced occurrence frequency measure to take care of categorial variable, as well as adopted a weighting strategy, instead of simply normalizing the data, that is driven by the information in both dataset. A variable-by-variable evaluation, in figure 7 is given to each component introduced. Although the overall performance is as expected, the error of one particular imputed variable, e18400, increased when complex models are used. It might because of the input features we chose for our models: the three features we used, wage, marital status and number of exemption, themselves do not have strong interpretabilities toward this imputed variable, and thus introducing models with higher complexity will make results worse-off. [7] Depending on the importance of this variable, we could improve the performance by developing another independent matching algorithm in the future.

With help of CEX dataset, data manipulations and improvements, model selections, and robustness test, we are able to obtain a more promising version of puf dataset with imputed itemized expenses. It would still be favorable if we could find some official benchmark and compare it with our results. Once available, another way to improve our imputation accuracy could be scoring and targeting at such benchmark.

References

[1] The Joint Committee on Taxation (2015), *Estimating Changes In The Federal Individual Income Tax: Description Of The Individual Tax Model*, JCX-75-15, The Joint Committee on Taxation.

[2] Kumcu, A. (2012), *No Longer Tax Exempt: Income Tax Calculation In The Consumer Expenditure Survey*, April Press, Monthly Labor Review, Bureau of Labor Statistics.

[3] Boriah, S.; Chandola, V.; Kumar, V. (2008), *Similarity measures for categorical data: A comparative evaluation*, SIAM, Proceedings of the 2008 SIAM International Conference on Data Mining.

[4] Hasti, T.; Tibshirani, R.; Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer Series in Statistics.

[5] Little, R.; Rubin, D. (2002), *Statistical Analysis with Missing Data*, 2nd Edition, Wiley Series in Probability and Statics.

[7] Recall our model evaluation is based on overall mean square error, instead of one particular variable, and thus it's likely that the overall MSE is decreasing while MSE of one variable is increasing.