Comparison of Random Forest and Multiple Imputation for Imputing Missing Data: a Case Study of the Education Panel Survey of the City of China

Li LI[1,2] Darran P. Goshawk[2]

1. School of Humanities and Social Science, Institute for Empirical Social Science Research, Xi'an Jiaotong University, Xi'an 710049,Shaanxi, China

2. Faculty of Architecture, Computing & Humanities, University of Greenwich, London SE10 9LS, UK

Abstract

In quantitative social research, choosing the most appropriate method for missing data imputation is a challenging work. In this work, 3 cutting-edge methods of data imputation, iterative imputation method based on a random forest (MissForest), multiple imputation by Chained Equations (MICE) and random forest-based MICE, were compared to find which method has the smallest imputation error. By using the normalized root mean squared error (NRMSE) and the proportion of falsely classified entries (PFC) as indicators, the results show that MissForrest is the most accurate method to handle missing data. The efficiency of the 3 methods were also investigated and compared. Implications for social research are suggested.

Keywords: Imputation, Random forest, CEPS

1. Introduction

In social surveys, missing data frequently occurs(Brunton- et al., 2014), this probably because some individuals forget/refuse to answer several specific questions (item non-response), or individuals refuse to interview although they were selected (unit non-response). As P. McKnight *et.al.* stated in their book <Missing Data: a Gentle Introduction>:(McKnight et al., 2007)

> *"As the old saying goes, the only certainties are death and taxes. We would like to add one more to that list: missing data."*

Generally speaking, missing values in social surveys seem inevitable. Simply drop the observations (listwise deletion or complete case analysis) with missing values will reduce the sample size and then makes the quantitative analysis unreliable. More importantly, missing data can also result in biased of the relationships of interest between independent and dependent variables. A widely cited item non-response example is: people on high incomes may be less willing to provide details of how much they earn in a survey (and hence missingness on income depends on a respondent's income). This means that if the observations with missing income data were simply removed, the result will necessarily be systematic biased(Brunton- et al., 2014; 刘凤芹, 2009). Therefore, methods other than listwise deletion should be used.

A common technique for handling item non-response is imputation.(Graham, 2009) Imputation is the process of replacing missing data with substituted values; hence a "complete" dataset is obtained and then can be analyzed with traditional analysis methods. A good imputation should not only give the best possible predictions of the missing values, more importantly, it should also reduce the selection bias associated with only using the "complete case analysis"(Andridge and Little, 2010; LITTLE and RUBIN, 1989). To achieve this, the objective of imputation is to exploit the statics information of population parameters, as the aim of social survey is making inferences about population quantities including regression coefficients. This requires a careful decision on which imputation method should be used. The main purpose of this paper is to help to address this issue, based on a city data of China.

The paper is organized as follows: in section 2, some of the widely used imputation methods will be briefly introduced and their pros and cons will be reviewed, especially for the state-of-the-art machine learning based imputation methods; In Section 3, the dataset and 3 imputation methods used in this work will be described and parameters for these method will be determined; In Section 4, the imputation result will be compared and discussed; Finally we will give concluding remarks in Section 5.

2. Literature review

Imputation theory is continuously developing and there are many theories, methods, models and approaches developed by scientists in different research area to account for missing data(Finch, 2010, 2015). Table 1 lists some of the widely used imputation methods. It should be noted that more and more new methods are being developed, e.g. the recent proposed fractional imputation method(2015), therefore attention to new information of this research field is required. Other data augment methods beyond imputation, such as expectation–maximization (EM) and maximum likelihood(Yuan et al., 2012), are not within the scope of the review and hence are not included in Table 1.

Table 1 List of imputation methods

Method	Brief description
Listwise deletion (Complete case analysis)	delete any observation that has missing data
Last observation carried forward(LOCF)	substituting missing measurements with the last observed measurement in Panel data
Mean substitution	the mean of the total sample for a variable is substituted for aU of the missing values in that variable
Hotdecking,	Identifies an observation with complete data who is similar to the observation with incomplete data, then uses that observation's data to replace the missing value.
Regression Imputation	variable with missing data is used as the dependent variable
KNN	For each missing value find the k nearest neighbours, then impute the missing value using the imputation function on the k-length vector of values found from the neighbors.
Multiple imputation by Chain Equations/by Fully conditional specification (MICE)	The missing values are imputed by multiple regression sequentially based on different types of missing covariates, and Gibbs sampling is used to estimate the parameters.
MissForest	Using a random forest trained on the observed values of a data matrix to predict the missing values.

From the practical point of view of quantitative social researchers, the most interesting thing is how to choose a best imputation method to give better quantitative analysis. There are a number of papers about comparisons of different imputation method in the literature, which are listed in Table 2 including the imputation methods compared and the type of dataset under investigation. In order to give a comprehensive view, some dataset other than social survey, such as medical, DNA and economics, are also included. Based on these paper, it's hard to say one method is superior the others, because imputation quality relates to the nature of the dataset's structure, which differ from one to another. Therefore, in order to find a suitable method for social survey data imputation, the characterises of social survey data should be discussed firstly.

Table 2 Literature review of imputation methods comparisons. The full names of imputation methods are listed in Table 1.

Dataset	Methods	Ref
Various	MissForest KNN MICE	(Stekhoven and Buhlmann, 2012)
Drug trial	LOCF, MICE MICE Listwise deletion	(Jorgensen et al., 2014)
Phenomic datasets	KNN, MICE Mean, missForest	(Liao et al., 2014)
Social survey	Mean substitution Listwise deletion Hotdecking Regression imputation	(Saunders et al., 2006)
Bio-data	KNN, MICE singular value decomposition etc.	(Mandel J, 2015)
Life-history trait datasets	KNN, MICE MissForest	(Penone et al., 2014)

As discussed by Schenker et.al(Schenker et al., 2006), survey data is complex and has the following characteristics: 1) Hierarchical. Variables are reported at different levels, e.g. family or person level. 2) Dependency. Variables structurally depend on each other, e.g. some families did not report exact income values but did report income categories, which should be used to form bounds for exact income. According to these features, multiple imputation (MI) has been considered a standard approach for general-purpose estimation under item non-response in survey sampling. MI was proposed by Rubin(RUBIN, 1976) to replace each of missing data with multiple plausible values to reflect the full uncertainty in the prediction of missing data. Furthermore, Multiple Imputation by Chained Equations (MICE) is a popular method being implemented in many statics software, e.g. Stata, SAS. The missing values are imputed by multiple regression sequentially based on different types of missing covariates, and Gibbs sampling is used to estimate the parameters. Although now MICE is a mean-stream method for missing data imputation in social survey data, fast developing machine learning based imputation methods, as listed in Table 1, offer an alternative choice for quantitative social researchers.

It should be noted that among these methods, random forest is a new, fast developing machine learning methods. Random Forests has a variety of advantages over other statistical and machine learning techniques. The method is robust to error in the data, does not have the same assumptions of normality, linearity, and is virtually immune to the effects of multi-colinarity. Due to these advantages, it has been widely used in many fields, the reason for its few usage in social science(Jones and Linder, 2014) probably is the lack of explanation power: it acts like a 'block box' and hence it is not so intuitive as regression. But for imputation it is good as only predictive power is needed. In recent year it begins to be applied to missing data imputation(Hapfelmeier, 2012; Pantanowitz and Marwala, 2009a, b). Compared to traditional regression based imputation method such as MICE, one advantage of random forest is that all interaction between the variable is considered, which makes the prediction of values more accurate. Therefore, a comparison with the main-stream MICE method is necessary. To our best knowledge, there hasn't such a research for social science survey dataset. It should be noted that a imputation method combining MI and random forest has also been developed(Shah et al., 2014), which is also include in this paper as a comparison. The technical description of these methods can be found in Section 3.

3. Method

3.1 **Imputation Algorithms**

All the 3 imputation methods were implemented in R. The 3 packages are: 1) missForest: Nonparametric Missing Value Imputation using Random Forest; 2) mice: Multivariate Imputation by Chained Equations; 3) CALIBERrfimpute: Multiple imputation using MICE and Random Forest. In the rest of this paper, we mention these 3 methods as "missForest", "MICE", "RF+MICE", respectively.

The mechanism and implementation of these 3 algorithms are briefly introduced in the following paragraphs.

Firstly we brief introduce Random Forest. Random Forests is a machine learning methodology by which decision trees are iteratively grown and used to classify instances of data. Schematically, Random Forests works by:

I. Selecting a subset of variables from the total set for use in classification as splitting criterion.

II. For N training samples, sample N cases at random with replacement

III. Grow each tree with the N cases previously sampled utilizing the CART methodology. a. The subset of variables at step I is used to split at each node with the best split based on node purity.

b. Grow each tree to the maximum size possible.

All the trees iteratively grown this way constitute a “forest” or “ensemble.” Each tree “votes” by classifying each instance. The modal classification across all of the trees is the final predicted category (Pantanwitz & Marwala, 2008).

MissForest, as an application of random forest for imputation, treats the variable of the missing value as the dependent variable and uses other related independent variables to impute the missing data. This is done by growing a random forest (the “model”) of classification and regression trees for the final prediction. The method is repeated until the imputed values reach convergence. The package CALIBERrfimpute is Functions to impute using Random Forest under Full Conditional Specifications (Multivariate Imputation by Chained Equations).

3.2 **Imputation performance assessment**

The imputation performance is assessed using the normalized root mean squared error [NRMSE, Oba et al. (2003)] for continuous variables

$$\mathrm{NRMSE} = \sqrt{\frac{\mathrm{mean}((\mathrm{X}_{\mathrm{true}} - \mathrm{X}_{\mathrm{imp}})^2)}{\mathrm{var}(\mathrm{X}_{\mathrm{true}})}}$$

where $\mathrm{X}_{\mathrm{true}}$ is the complete data matrix and $\mathrm{X}_{\mathrm{imp}}$ the imputed data matrix. For categorical variables, proportion of falsely classified entries (PFC) is used for the categorical and binary/logical missing values. For both type of variables, smaller value means better imputation performance, i.e. a good imputation gives a value close to 0 and a bad imputation gives a value around 1.

The dataset containing missing data is generated in the following ways: firstly, the complete dataset (CD) were generated from the original raw dataset (RD) with missing values. Secondly, missing values were introduced by randomly deleted 10%, 20% or 30% observations of the dependent variable to obtain the dataset with missing values (MD). Then the 3 imputation methods were applied to MD, respectively. Finally, the performance was assessed by calculating the PFC for binary/logical variables and NRMSE for continuous variables, according to the imputed and the real values. The NRMSE and the PFC were estimated on 20 randomly generated MDs from the same CD.

It should be noted that out-of-bag (OOB) error estimate for the imputed variable can be obtained when a random forest is fit to the observed part of a variable. OOB error, also called OOB estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating to sub-sample data sampled used for training. D.J. Stekhoven et.al. proposed that OOB error is a good indicator of the 'true imputation error'. As in this work we can get the true imputation error, so OOB errors are not shown in this work.

3.3 Introduction to dataset CEPS

The China Education Panel Survey (CEPS) is conducted by National Survey Research Center (NSRC) at Renmin University of China. It is a large-scale, nationally representative, longitudinal survey starting with two cohorts – the 7th and 9th graders in the 2013-2014 academic year. It is focused on students progress through various educational stages. This data contains not only the students educational information ,but also multiple contexts of families, school processes, communities and social structure. The CEPS applies a stratified, multistage sampling design with probability proportional to size (PPS), randomly selecting a school-based, nationally representative sample of approximately 20,000 students in 438 classrooms of 112 schools in 28 county-level units in mainland China. The CEPS data is timely and significant because it captures the educational development during the rapid social change in China, providing rich and invaluable data source for researches in social sciences, policy makers, and school administrators.

Due to the great difference between urban and rural in China and the shadow education is much popular in urban, the quantitative data reported in this paper were derived from city sampling. The sample number was 9,487.

3.4 **Variables: literature based and Missing data analysis**

Shadow education was chosen as the research topic to be investigated. As an important mechanism of social mobility and social reproduction, education contains school education and shadow education. With the support of the policy of nine - year free compulsory education and the policy of balanced development of compulsory education, the school education is governed more and more, which partly reduced the educational inequality. While the shadow education expanded at the same time, and students especially the students in city participated in many kinds of shadow education to improve their educational scores. The competition of education has changed from school to outside. However, the inequality of shadow educational opportunity hasn't been considered enough, neither the policy makers nor the theory. Based on these, we focused on the shadow education in middle school of the city of China.

3 dependent variables were used: 1) Whether a student participated in shadow education, 2) the types of shadow education the child participated in, and 3) how much money the parents paid for the shadow education for one school term, depends on many independent variables. Then we found independent variables following a two-step pick-up. Firstly we chose relative independent variable based on existed researches, then verify the relationship by random forest importance measurement.

Then the random forest importance measurement was performed. Variable importance is one important function of random forest. Rather than characterization of the dependence of independent variable(s) on the dependent variable, variable importance describes how the model's ability to predict dependent variable depends on a particular independent variable. Fig.1(a) shows the variable importance of independent variables listed in Table 3. Here only one dependent variable "whether participated in shadow education" is shown as an example. Larger value means the independent variable is more "important" to the dependent variable. That is, if xj is strongly related to y, then permuting its values will produce a systematic decrease in the model's ability to predict y, and the stronger the relationship between xj and y, the larger this decrease. From Fig.1a, it can be seen most independent variables strongly correlates to the dependent variable. However, b32 and ba12 make neglect contribution to the dependent variable, therefore these two variables will not be considered in the following study.

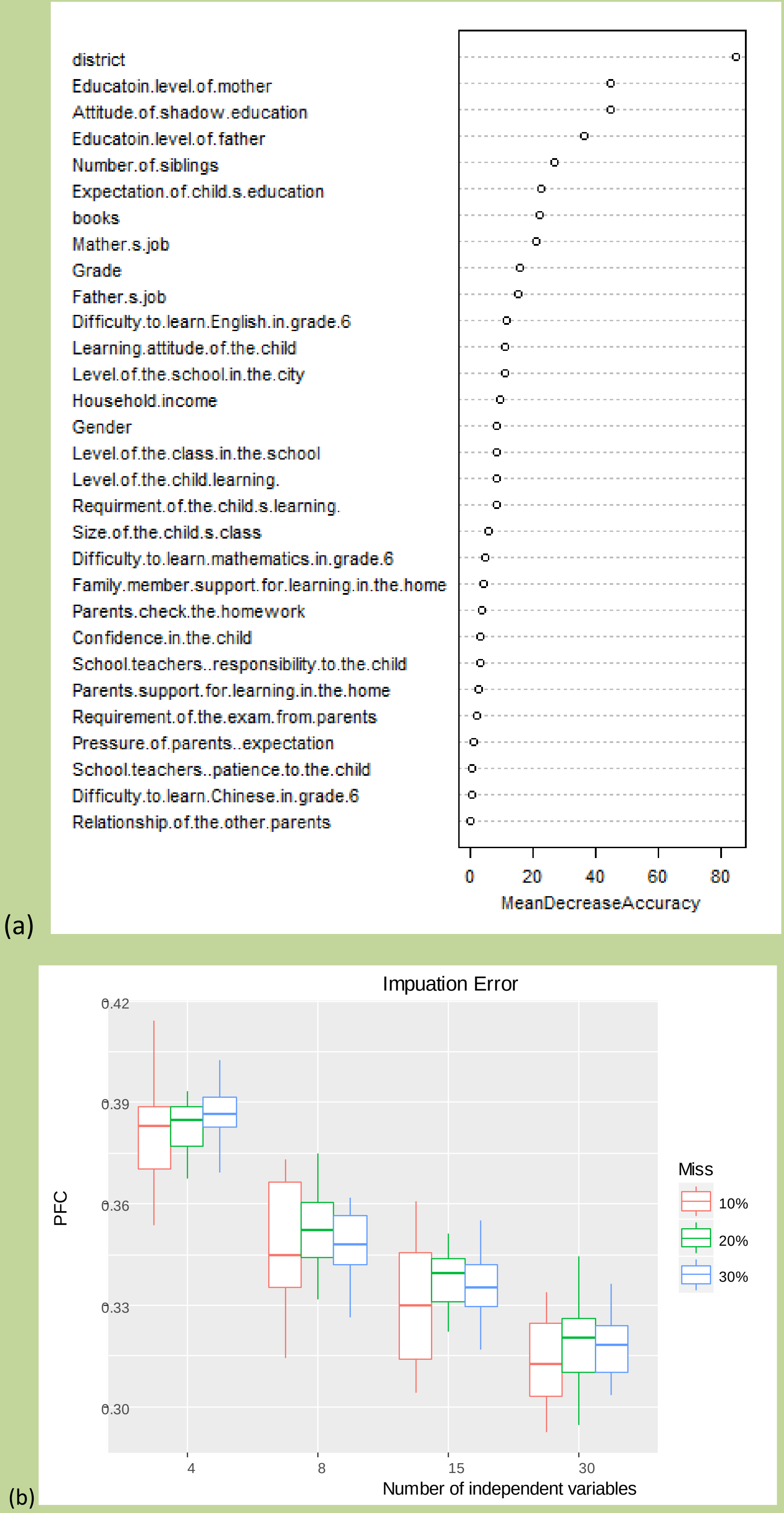

Fig.1 (a) variable important of independent variables.(b) the influence of number of independent variables on the imputation performance

Based on the ascending order of independent variable's importance as shown in Fig.1a, we can investigate the influence of number of independent variable on the imputation performance. We take the first 4, 8 and 15 most important variables, as well as all the variables, to impute the dependent variable, "whether attend shadow education". The results are shown in Fig.1(b). Two trends can be cleanly seen: 1) with increasing number of independent variables, the imputation error of dependent variable decreases. When 4 most related variables are used, the imputation error is 0.37~0.38; When the number of independent variables increases to 30, the imputation error is reduced to 0.31~0.32. 2) the imputation error for low missing rate data is slightly lower than a higher missing rate data, which can be explained by the nature of random forest data imputation: the more data used, the more accurate the imputation.

4. Results

Figure 2 gives the results of comparison of 3 imputation methods on 10%, 20% and 30% missing proportion, and for binary/logical, continuous and category variables, respectively. The results were computed from 20 randomly generating dataset as described in section 3.2. It can be seen in all 3 types of variables, missForest has the best performance, and reduces imputation error in many cases by ~30%. The performance of MICE and RF+MICE is similar, both much worse than missForest. The reason for the bad performance of RF+MICE probably is that it's not based on a iteration method.

As for the performance, the binary/logical variable is acceptable, where missForest gives an error around 30%. But for continuous variable, the performance is bad and the imputation error is around 1, which will introduce systemic bias in the following analysis. The reason is probably the independent variables are mostly category/binary, while the dependent variable is continuous.

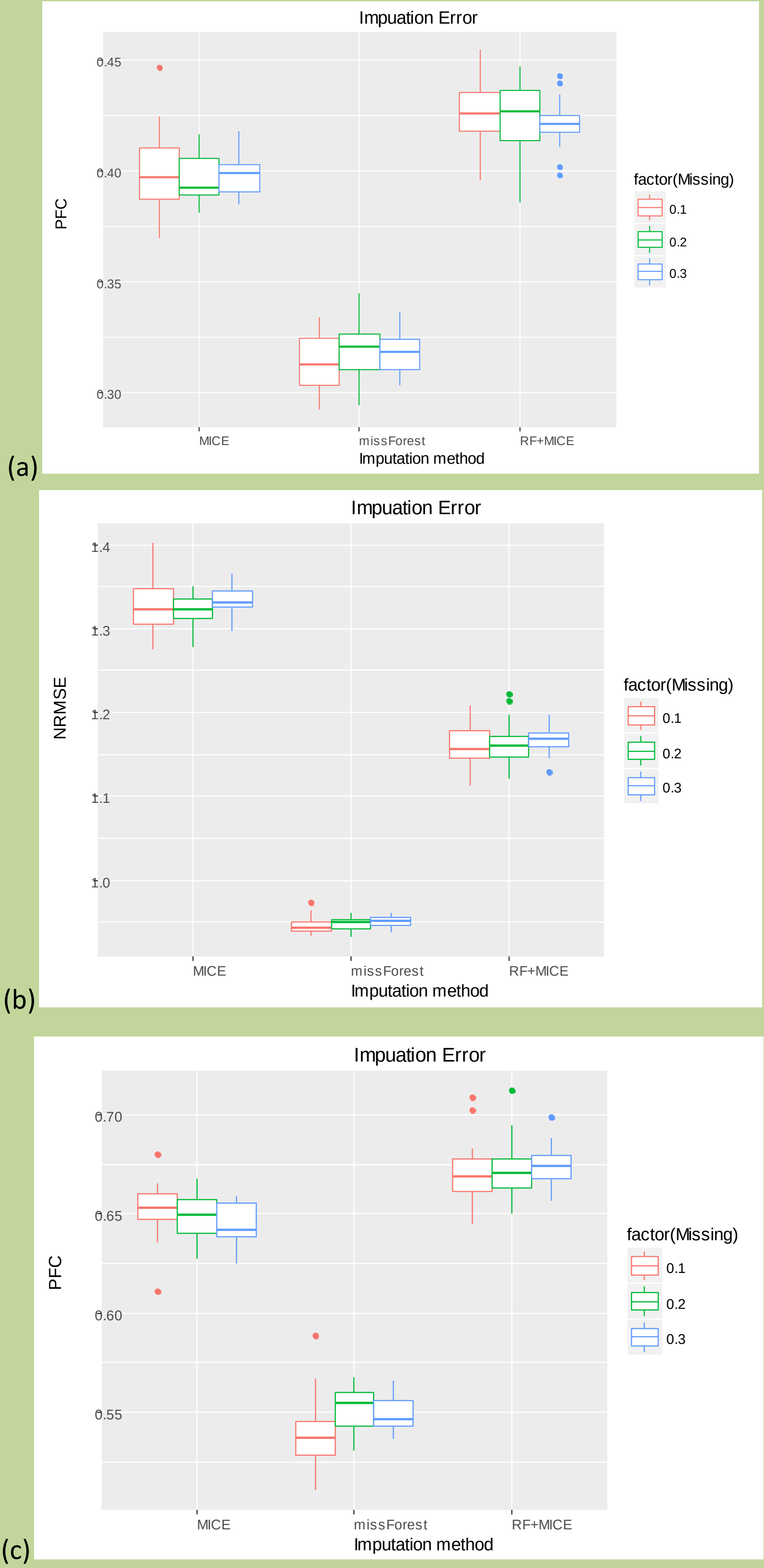

Fig.2 Imputation performance of a) binary/logical variable; b) continuous variable; c) category variable

We also compared the average running time of different methods. The propitiation of missing data is 10% for dependent variable. The running time is on 1 core in a 2 x 6 Core Intel Xeon E5645 computing node on Midplus (See acknowledgement). It can be seen that in all three case RF+MICE is the most fast. missForest uses less time for logical(binary) and category variable, while MICE uses less time for continuous variables.

Table 3 Average running time of the imputation methods Unit: second

	Logical	Continuous	Category
missForest	30s	420s	50s
RF+MICE	10s	13s	10s
MICE	60s	20s	180s

5. Concluding remarks

3 cutting-edge imputation methods were compared on CEPS dataset. Imputation based on random forest, MissForest, gives much better imputation performance than both MICE and RF+MICE, although the running time for continuous variables is much longer than the other two methods. It should also be noted that, as been pointed out for many times, the easiest way to handle missing data is to avoid missing values during the data collection process. Although this seems unrealistic considering the nature of the social science survey, the researcher should pay more attention during data collection. Furthermore, Based on the literature review and the results of this study, we propose that new methods should be continuously developed and implemented.

Acknowledgement

This work is financially supported by the Projects of Students of the Zheng Hangsheng Society Development Fund under grant No. 13ZHFD08 and the Key Projects of the National Social Science Fund of China under grant No. 14ASH001. L Li thanks to the support of PhD Mobility Program of Xi’an Jiaotong University.

Reference

2015. Fractional Imputation in Survey Sampling: A Comparative Review.

Andridge, R.R., Little, R.J., 2010. A Review of Hot Deck Imputation for Survey Non-response. Int Stat Rev 78, 40-64.

Brunton-, I., Smith, James, Carpenter, Mike, Kenward, Tarling, R., 2014. Multiple Imputation for handling missing data in social research.

Finch, W.H., 2010. Imputation Methods for Missing Categorical Questionnaire
Data: A Comparison of Approaches.

Finch, W.H., 2015. Missing Data and Multiple Imputation in the Context of Multivariate Analysis of Variance. The Journal of Experimental Education, 1-17.

Graham, J.W., 2009. Missing data analysis: making it work in the real world. Annu Rev Psychol 60, 549-576.

Hapfelmeier, A., 2012. Analysis of Missing Data with Random Forests.

Jones, Z., Linder, F., 2014. Random Forests for the Social Sciences.

Jorgensen, A.W., Lundstrom, L.H., Wetterslev, J., Astrup, A., Gotzsche, P.C., 2014. Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial. PLoS One 9, e111964.

Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F.C., Tseng, G.C., 2014. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinformatics 15, 346.

LITTLE, R.J.A., RUBIN, D.B., 1989. The Analysis of Social Science Data with Missing Values. Sociological Methods & Research 18, 292-326.

Mandel J, S.P., 2015. A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics & Biostatistics 06.

McKnight, P.E., McKnight, K.M., Sidani, S., Figueredo, A.J., 2007. Missing Data: A Gentle Introduction.

Pantanowitz, A., Marwala, T., 2009a. Evaluating the Impact of Missing Data Imputation
through the use of the Random Forest Algorithm.

Pantanowitz, A., Marwala, T., 2009b. Missing Data Imputation Through the Use of the Random Forest Algorithm. 116, 53-62.

Penone, C., Davidson, A.D., Shoemaker, K.T., Di Marco, M., Rondinini, C., Brooks, T.M., Young, B.E., Graham, C.H., Costa, G.C., Freckleton, R., 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods in Ecology and Evolution 5, 961-970.

RUBIN, D.B., 1976. Inference and missing data. Biometrika 63, 581-592.

Saunders, J.A., Morrow-Howell, N., Spitznagel, E., Dori, P., Proctor, E.K., Pescarino, R., 2006. Imputing Missing Data: A Comparison of Methods for Social Work Researchers.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., Cohen, A.J., 2006. Multiple Imputation of Missing Income Data in the National Health Interview Survey. Journal of the American Statistical Association 101, 924-933.

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol 179, 764-774.

Stekhoven, D.J., Buhlmann, P., 2012. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112-118.

Yuan, K.H., Yang-Wallentin, F., Bentler, P.M., 2012. ML versus MI for Missing Data with Violation of Distribution Conditions. Sociol Methods Res 41, 598-629.

刘凤芹, 2009. 基于链式方程的收入变量缺失值的多重插补. 统计研究, 71-77.