Journal of Risk & Control, 2021, 8 (1), 1-18
ISSN: 2056-3701 (print version), 2056-371X (online)
https://doi.org/10.47260/jrc/811
Scientific Press International Limited

Multiple Imputation for Missing Values with an Empirical Application

Theodora Sotiropoulou[1], Stefanos Giakoumatos[2]
and Antonios Georgopoulos[3]

Abstract

Missing data are the most common problem in many research areas. For cross-section and time-series data, imputation can be a challenging problem. The most widely used method for filling missing observations is the multiple imputation which increase the number of the available data and thereby reducing biases that may occur when observations with missing values are simply deleted. The main purpose of this paper is to employ a bootstrapping expectation–maximization (EM) algorithm in order to impute missing values mainly to economic data. In the application we use a dataset that is consisted by annual panel data for the 27 countries of the European Union covering the period 2000-2017. The data were obtained from the databases of World Bank and Eurostat namely the Global Financial Development Database, The Standardized World Income Inequality Database by Solt (2019) and the World Development Indicators. Different indicators were chosen representing the development of banking system and stock markets, economic growth, economic inequality, innovation, fiscal policy, physical and human capital, and trade openness. Finally, diagnostic tools are used inspecting the imputations that are created.

Keywords: Multiple imputation, Amelia II, Economic data, Financial development, Inequality.

[1] PhD candidate, Dep. of Business Administration, University of Patras, Greece.
[2] Professor, Dep. of Accounting and Finance, University of Peloponnese, Greece.
[3] Professor, Dep. of Business Administration, University of Patras, Greece.

Article Info: *Received:* May 10, 2021. *Revised*: May 20, 2021.
Published online: June 21, 2021.

1. Introduction

Missing data is a conventional problem in social science, since data are incomplete due to respondents do not answer every question or countries do not collect statistics every year. Most statistical methods assume the absence of missing data and presume complete information for all the variables included in the analysis. For researchers, missing data are a nuisance because, even if the missingness is not the main focus of their research, they have to address this issue as most data analysis procedures are designed for complete data matrices (Schafer & Graham, 2002). A relatively few absent observations on some variables can dramatically shrink the sample size, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. Appropriately dealing with missing can be challenging as it requires a careful examination of the data to identify the type and pattern of missingness, and also a clear understanding of how the different imputation methods work.

A growing literature suggests how researchers deal with missing data can affect model estimates and standard errors (Schafer, 1997; Vriens and Melton, 2002; Schafer and Graham, 2002; Raghunathan, 2004; Little and Rubin, 2002). The most common method with easy application is the listwise deletion which use only those cases with complete information. An alternative method is the complete-case analysis filling the missing observations with the mean of the observed cases on that variable. More recently, statisticians have advocated methods that are based on distributional models for the data such as maximum likelihood and multiple imputation (Little, 1992; Little & Rubin, 1987; Schafer, 1997). The missing values should be replaced with rational values, to carry out an analysis based on a "complete" dataset. This approach of handling missing values is called Imputation. Imputation is an immense field of study, where a lot of research has already taken place. Popular techniques are Multiple Imputation (Rubin, 1987), Expectation-Maximization (Dempster et al., 1977), Nearest Neighbor (Vacek and Ashikaga, 1980) and Hot Deck methods (Ford, 1980). The method of imputation assumes that the estimation of missing values derives from a predictive distribution which is based on observed values.

Many sources of economic data cover only a limited set of countries and time periods. The main scope of this study is the application of multiple imputation method using the statistical package Amelia II of the software R, in the context of time series cross section data. The dataset consists of variables measured the financial development, economic growth, income inequality and other economic indicators for the 27 countries of the European Union covering the period 2000-2017.

The remaining sections of this study are organized as follows: section two offers an introduction to multiple imputation and discuss the algorithm and the package used in this study. Section three describes the dataset and the application results are presented. Finally, section four provides the conclusions.

2. Theory of multiple imputation

Missing observations make it difficult for analysts to realize the data analysis. Types of problems that are usually associated with missing values are the loss of efficiency, complications in handling and analyzing the data, bias resulting from differences between missing and complete data and reduction of statistical power. Decision on selecting an appropriate method for handling missing observations on time series, cross-section, time-series cross-section data depends on the missing data pattern and on the missing-data mechanism.

A missing data pattern refers to the configuration of observed and missing values in a data set, whereas missing data mechanisms describe possible relationships between measured variables and the probability of missing data. Thus, a missing data pattern simply describes the location of the "holes" in the data and does not explain why the data are missing.

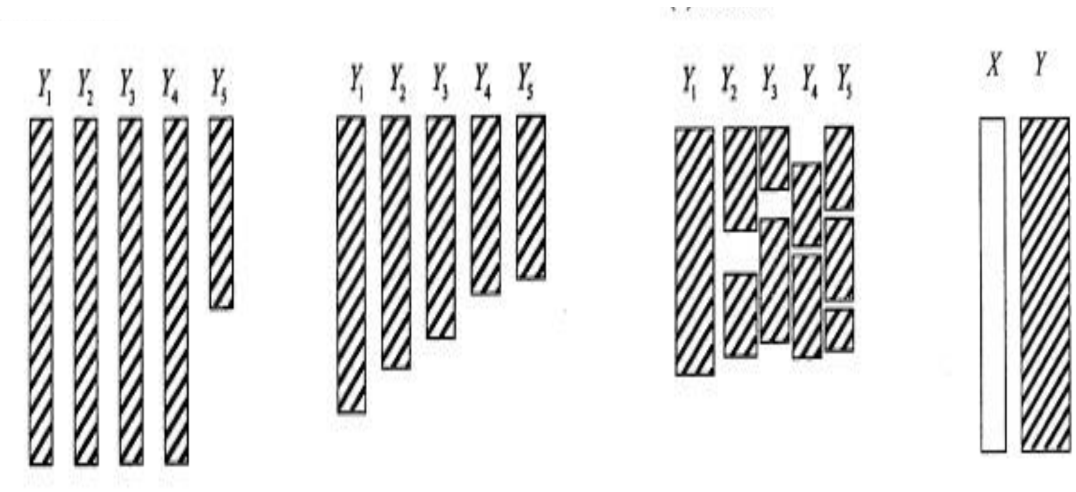
Y_1 Y_2 Y_3 Y_4 Y_5
Y_1 Y_2 Y_3 Y_4 Y_5
Y_1 Y_2 Y_3 Y_4 Y_5
X Y

Figure 1: Missing data patterns

Various missing data patterns are tracked in the missing data literature. Figure 1 depicts different types of missing data patterns. Particularly, on the left side, the univariate pattern has missing values isolated to a single variable. The second pattern is a monotone missing data pattern and typically associated with a longitudinal study where participants drop out and never return. Visually, the monotone pattern resembles a staircase, such that the cases with missing data on a particular assessment are always missing subsequent measurements. The third type is the general or arbitrary missing data pattern and perhaps the most common configuration of missing values. A general pattern has missing values dispersed throughout the data matrix in a haphazard fashion. Finally, the latent variable pattern is unique because the values of the latent variables are missing for the entire sample.

Missing data mechanisms play a significant role in missing data theory. Rubin (1976) proposed a classification system for missing data.

- Data are missing completely at random (MCAR) if the probability of missing data is unrelated to any information in the dataset
- Data are missing at random (MAR) if missing data depends only on the observations that are observed.
- Data are missing not at random (MNAR) when the probability of missing data depends on both observed and missing values.

The challenge of missing data requires to identify the pattern and the mechanism of missingness in order to determine which method can be used to deal with missing data.

Multiple Imputation is a powerful statistical technique developed by Rubin in the 1970s for analyzing datasets containing missing values. Multiple imputation is a method for estimating missing information generating several estimates for each missing value, using the variation between estimates as a measure of the uncertainty associated with imputation. The imputation approach makes two assumptions that the data are missing at random (MAR) and the complete data, both observed and unobserved has multivariate normal distribution.

It is a Monte Carlo technique that requires three distinct steps: the imputation phase, the analysis phase, and the pooling phase.

- The imputation phase creates multiple copies of the data set (e.g., m = 3), each of which contains different estimates of the missing values.
- The analysis phase is to analyze the filled-in data sets. This step applies the same statistical procedures that you have used if the data had been completed. Procedurally, the only difference is that you perform each analysis m times, once for each imputed data set. The analysis phase yields m sets of parameter estimates and standard errors.
- The pooling phase is to combine everything into a single set of results. According "Rubin combination rules" (Rubin 1987), the pooled point estimate is equal to the average of the m separate estimates, while its variance is equal to a weighted sum of the estimated variances within and between the m datasets.

The three-step process (imputation, analysis, pooling) is common to all multiple imputation procedures, a variety of algorithms for the imputation phase has been proposed (King, Honaker, Joseph, & Scheve, 2001; Lavori, Dawson, & Shera, 1995; Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; Royston, 2005; Schafer, 1997, 2001; van Buuren, 2007). These algorithms address different types of problems.

Various tools are available for performing multiple imputation (STATA, SAS, SPlus, MICE). Amelia II program in R developed by Honaker, King, and Blackwell (2011) is the most widely used multiple imputation software in social and economic science. Amelia II is the successor to the original Amelia program developed by King et al. (2001), implements joint multivariate normal multiple imputation and employs a bootstrapping expectation–maximization (EM) algorithm to take draws from the posterior distribution. Amelia II simply creates and implements an imputation model, generates imputed datasets replacing the missing values, and checks its fit using diagnostics. The Amelia II program has several improved capabilities such as the imputation of many more variables, with many more observations, in much less time, great simplicity and power, never crashes and is much faster than the alternatives. Amelia II provides valid and more accurate imputations for cross sectional, time-series, and time-series-cross-section data and many diagnostic functions to check the validity of imputation model.

3. Application

One of the most important issues in literature is the key drivers of economic growth. Financial development plays a vital role on economic growth from several theoretical and empirical perspectives. Additionally, the recently rising inequality in the world has re-highlighted the debate about inequality-growth nexus. The investigation of the relationships between financial development, economic growth, income inequality and other economic variables, is still a matter of discussion. Many sources of economic data cover only a limited set of countries over certain time periods resulting in missing data problems. Most methods of statistical analysis require complete data for all variables since missingness can dramatically shrink the sample size, the statistical power weakens and the parameter estimates may be biased.

3.1 Data

The dataset consists of annual panel data for the 27 countries of the European Union covering the period 2000-2017. The data for the financial development were acquired from the Global Financial Development Database, the measures of income inequality were come from The Standardized World Income Inequality Database by Solt (2019) and Eurostat, the indicators for innovation were obtained from the World Development Indicators and Eurostat and the data for economic growth and the control variables were obtained from the World Development Indicators.

Financial development does not have a direct measure. Thus, different indicators were chosen representing the characteristics both the banking system development and stock markets development.

Particularly, the indicators of financial development are the following:

- Private credit is defined as the credit provided by all financial intermediaries, excluding central banks, to the private sector as a percent of GDP.
- Liquid liabilities which are equal the ratio of liquid liabilities to GDP.
- Bank assets are the ratio of deposit money bank assets divided by deposit money bank assets and central bank assets.
- Domestic credit is equal to the domestic credit to private sector as a percent of GDP.
- Market capitalization is the total value of all shares in the stock market as a percent of GDP.
- Value traded is the value of all shares traded in the stock market as a percent of GDP.
- Interest margin measured by the rate of bank net interest margin which is the accounting value of bank's net interest revenue as a share of its average interest-bearing (total earning) assets.
- Turnover ratio expressed by the value of the traded shares in the domestic stock market divided by the total value of shares in the market.
- Z-score captures the probability of default of a country's commercial banking system.
- Non-performing loans is the ratio of non-performing loans to gross loans.
- Stock price volatility is defined as the average of the 360-day volatility of the national stock market index.

Economic growth is measured by:

- Growth rate of GDP per capita as an annual percentage.
- GDP per capita constant 2010 U.S. dollars.

Income inequality is expressed by:

- Gini coefficient is the income inequality estimates are based on Gini indices of disposable (post-tax, post-transfer) income.
- Income share is measured by Income quintile share ratio S80/S20 for disposable income by sex and age group.

Innovation is measured by:

- R&D expenditures are gross domestic expenditures on research and development (R&D), expressed as a percent of GDP.
- Patent is the number of patent applications to the European Patent Office (EPO) by priority year.

Control variables

- Investment is measured as the gross capital formation as a percent of GDP.
- Inflation is the annual change of consumer prices.
- Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product.
- Government expenditure is the General government final consumption expenditure as a percent of GDP.
- Human capital is measured by the Gross enrollment ratio for secondary school.

3.2 Results

When performing multiple imputation, the first step is to identify which variables to include in the imputation model. It is crucial to include at least as much information as will be used in the analysis model. That is, any variable that will be in the analysis model should also be in the imputation model. This includes any transformations or interactions of variables that will appear in the analysis model. It is often useful to add more information to the imputation model than will be present when the analysis is run. Since imputation is predictive, any variables that would increase predictive power should be included in the model, even if including them in the analysis model would produce bias in estimating a causal effect or collinearity would preclude determining which variable had a relationship with the dependent variable. A data set is constructed to analyse the relationships between financial development, economic growth, income inequality and other economic variables in European Union countries during the period 2000-2017. In fact, multiple imputation programs make no distinction between independent and dependent variables and only require to specify a set of input variables.

Secondly, the proportion of missing values for each variable is computed. Checking the summary statistics of the data, the Table 1 displays the missingness on many of the variables. These missing values in the dataset, reduce the sample size of 486 observations.

Table 1: Descriptive statistics of variables.

Variable	Mean	St. dev.	Min	Max	Observations	Missing
GROWTH	2.2767	3.7563	-14.2687	23.9855	486	-
lnGDP	10.1297	0.6959	8.2828	11.6260	486	-
lnPRIV	4.2473	0.5855	1.8550	5.5633	482	4
lnASSET	4.5892	0.0756	4.1632	4.9733	457	29
lnLLY	4.3367	0.6392	1.4260	6.8445	480	6
lnDOMCR	4.2681	0.6395	-1.6827	5.5344	486	
lnSMCAP	3.5112	0.9390	-0.3047	5.5101	431	55
lnVALTRADED	2.0354	1.9344	-3.5914	5.4928	427	59
lnINTEREST	0.6558	0.6637	-2.0748	2.2933	486	-
lnTURNOVER	3.1312	1.5897	-2.0337	5.8906	421	65
lnZSCORE	2.2715	0.7476	-4.0923	3.8623	485	1
lnNPL	1.2865	1.1461	-2.3026	3.8852	440	46
lnPRVOL	2.9774	0.4007	1.8467	4.1163	452	34
lnGINI	3.3620	0.1188	3.1091	3.5724	485	1
lnINCSHARE	1.5480	0.2307	1.1756	2.1187	362	124
lnRESDEV	0.1611	0.6498	-1.4832	1.3630	477	9
lnPATENT	5.3666	2.3302	-1.1087	10.1022	483	3
lnGOVEXP	2.9746	0.1425	2.4835	3.3299	486	-
lnINVEST	3.1225	0.1924	2.3240	3.7245	486	-
lnINFLATION	1.3520	0.9341	-2.2046	4.5146	486	-
lnTRADE	4.6478	0.4660	3.8159	6.0122	486	-
lnEDUSEC	4.6584	0.1355	4.3792	5.0995	478	8

The variables with missing values are private credit, bank assets, liquid liabilities, stock market capitalization, value traded, turnover ratio, Z-score, non-performing loans, stock price volatility, Gini coefficient, income share, R&D expenditures, number of patents and school enrolment. All variables have been transformed into their natural logarithms, except GROWTH which is not transformed and inflation is transformed as $\log\left(inflation + \sqrt{inflation^2 + 1}\right)$.

The third step is the evaluation of different missing value patterns in the data. One useful tool for exploring the missingness in a dataset is a missingness map. This is a map that visualizes the dataset in a grid and colors the data missingness. The columns of the grid are the variables and the rows are the cases. The missingness map allows for understanding the patterns of missingness in the data and can often indicates potential ways to improve the imputation model or data collection process. Figure 2 shows the pattern of missingness is arbitrary.

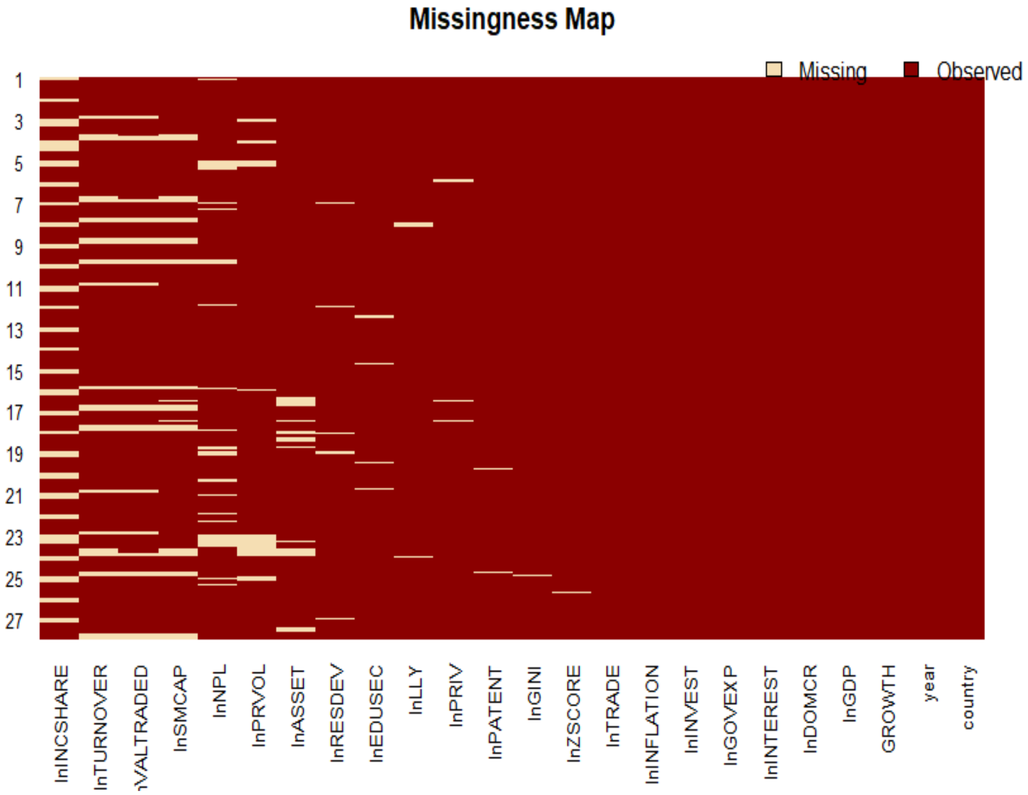

Figure 2: Missing map of dataset. Missing values are in tan and observed values are in red

Multiple imputation to handle missing data involves 3 steps:

a) Running an imputation model defined by the chosen variables to create imputed data sets. The missing values are filled in m=5 times to generate m=5 complete data sets since the literature of multiple imputation recommends between three and five imputed data sets (Rubin, 1987, 1996; Schafer, 1997; Schafer & Olsen, 1998).

b) The m=5 complete data sets are analyzed by using standard procedures.

c) The parameter estimates from each imputed data set are combined to get a final set of parameter estimates. According to Rubin’s (1987) procedure for combining parameter estimates, the multiple imputation point estimate is the arithmetic average of the m sets of estimates.

3.3 Imputation diagnostics

Imputed values need to be checked for their plausibility. Amelia II package provides several tools for diagnostics of imputation.

The Comparing densities is one check on the plausibility of the imputation model and tests the distribution of imputed values to the distribution of observed values. Obviously these distribution will not be identical as the missing values may differ systematically from the observed value. The distribution of mean imputations is represented by the red line while the distribution of observed values are represented by the black line for each variable. When, there are no missing values, the distribution of observed values is simply plotted with a blue line. Imputed values of income shares, non-performing loans and stock price volatility are very similar to observed values, however, the imputation of bank assets, liquid liabilities, stock market capitalization, value traded, turnover ratio, are quite different from the observed data.

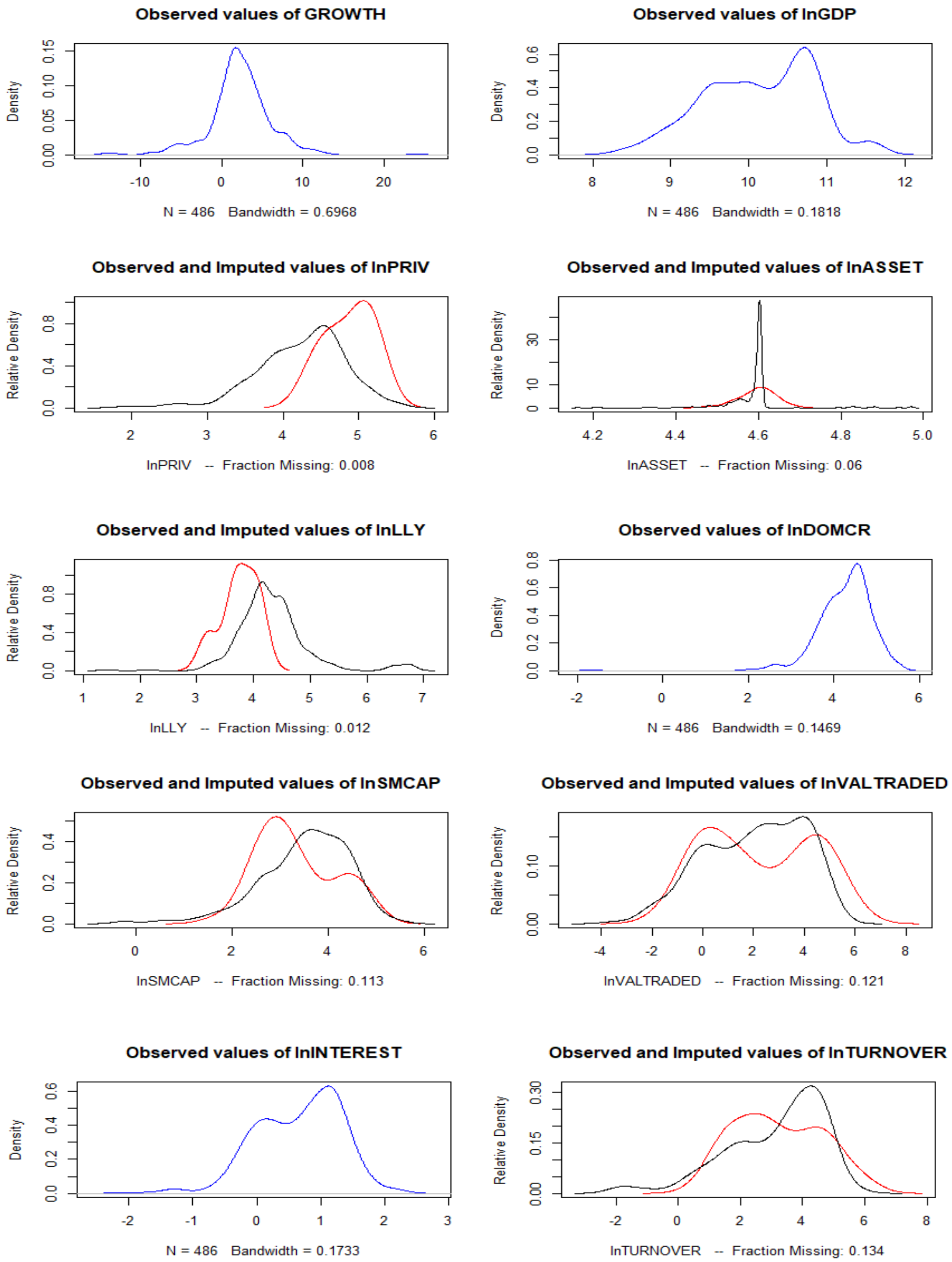

Figure 3a: The plots depict the distribution of mean imputations (in red) and the distribution of observed values (in black) for each variable with missing values.

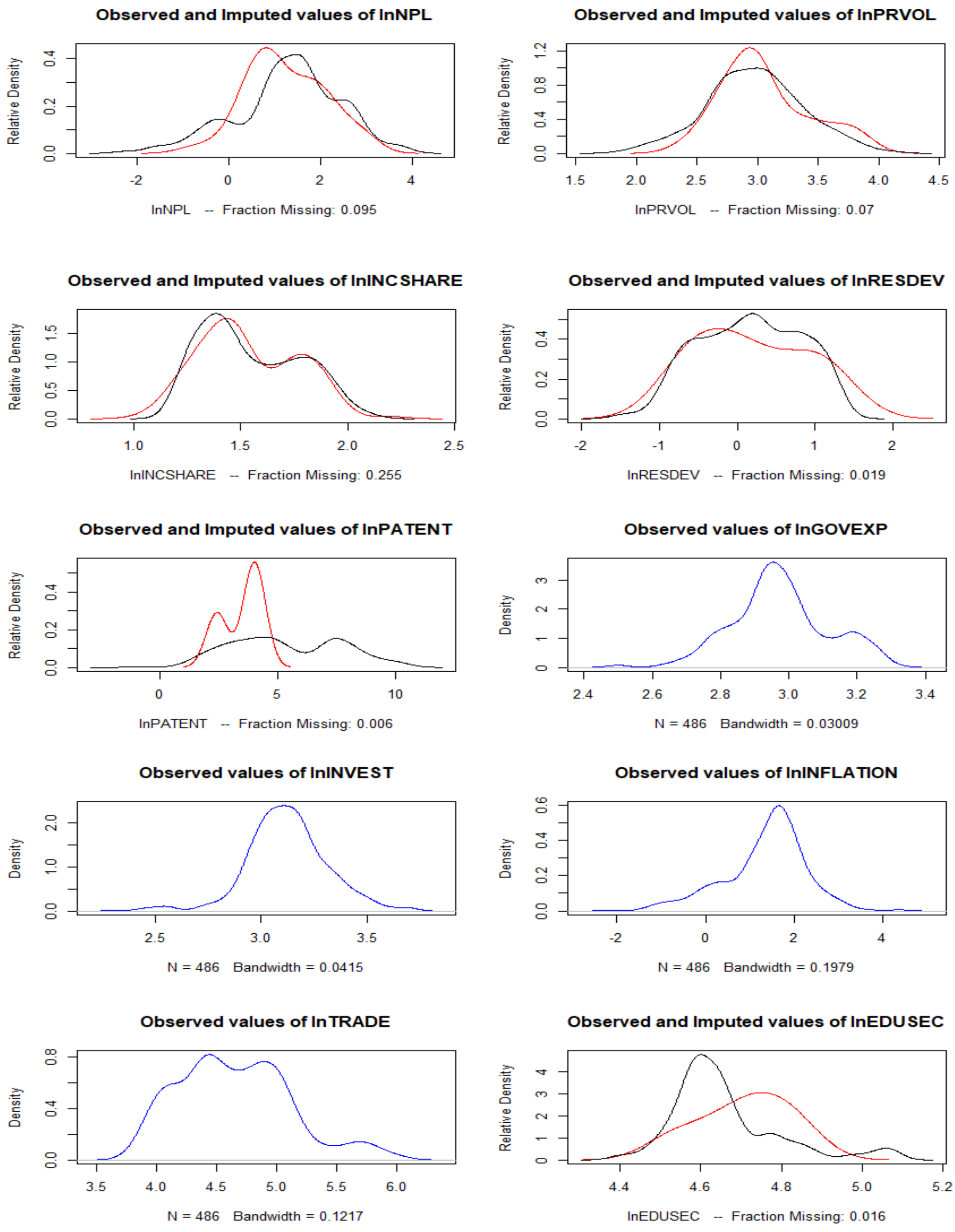

Figure 3b: The plots depict the distribution of mean imputations (in red) and the distribution of observed values (in black) for each variable with missing values.

Over-imputation is a technique to test the fit of the imputation model. It is impossible to tell whether the mean prediction of the imputation model is close to the unobserved value that is trying to be recovered. The horizontal line displays the actual observed value, and the vertical line denotes the imputed value pretending that the observed values are missing. The dots are mean value of imputation and lines represent the 90% confidence interval. If mean values are on the diagonal line, the imputation model is exactly accurate. It is obvious from Figure 4 that imputation model predicts well for missing values.

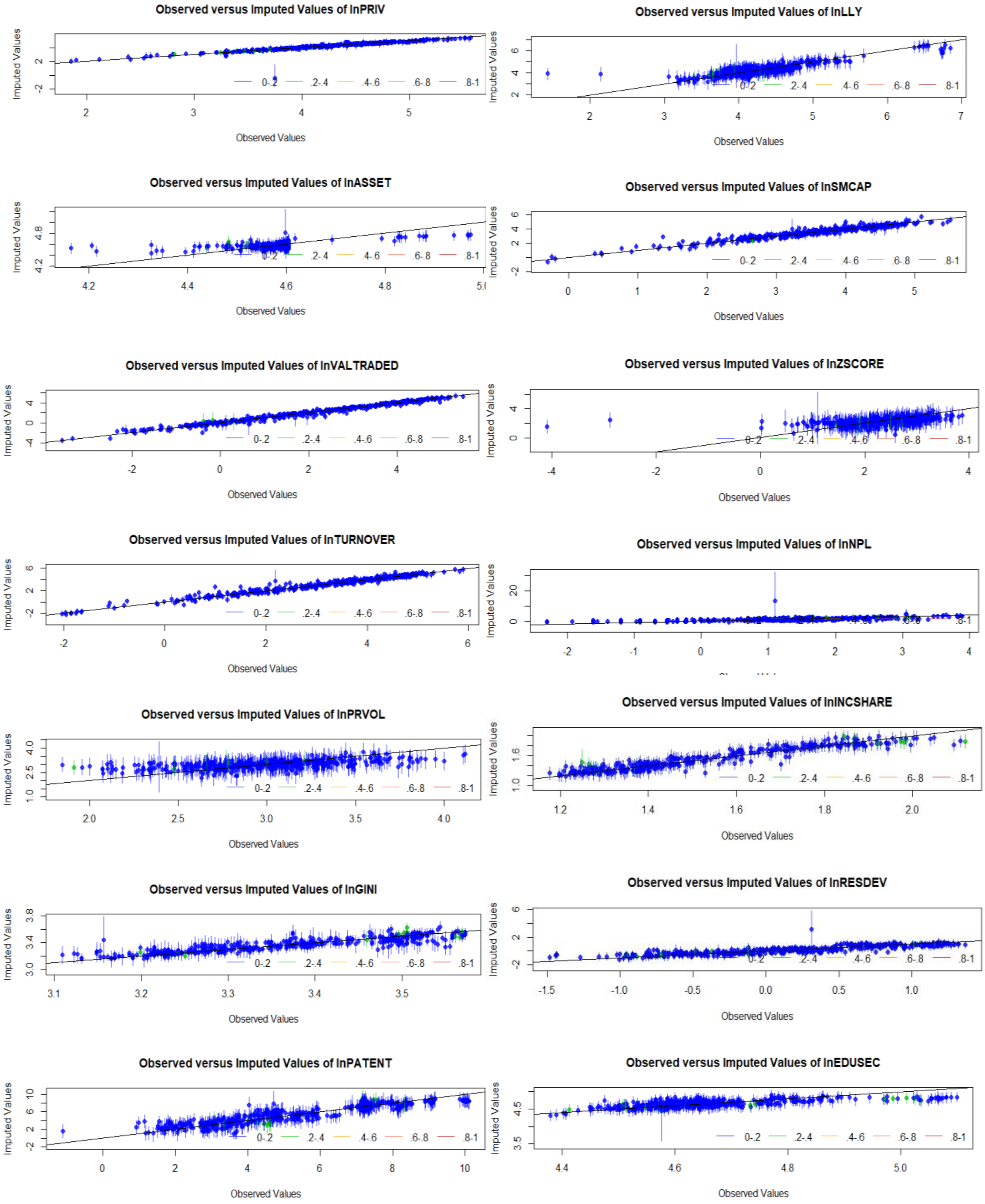

Figure 4: Over-imputation diagnostic graph for each imputed variables. The dots represent the mean imputation. Around ninety percent of these confidence intervals contain the y = x line, which means that the true observed value falls within this range.

The overdispersion is a diagnostic to run the EM chain from multiple starting values that are overdispersed from the estimated maximum and will display a graph of the paths of each chain. Since these chains move through spaces that are in an extremely high number of dimensions and cannot be graphically displayed, the diagnostic reduces the dimensionality of the EM paths by showing the paths relative to the largest principle components of the final mode(s) that are reached. On the left, the y-axis represents movement in the (very high dimensional) parameter space, and the x-axis represents the iteration number of the chain. Once the diagnostic draws the graph, the results show that all chains convergence to the same point. In other words, if all of the lines converge to the same point, then it is confident that starting values are not affecting the EM algorithm. On the right, the parameter space are visualized in two dimensions using the first two principal components of the end points of the EM chains. The iteration number is no longer represented on the y-axis, although the distance between iterations is marked by the distance between arrowheads on each chain. In one dimension, the diagnostic plots movement of the chain on the y-axis and time, in the form of the iteration number, on the x-axis. Figure 5 shows a well behaved likelihood, as the starting values all converge to the same point. The black horizontal line is the point where Amelia II converges when it uses the default method for choosing the starting values. The diagnostic takes the end point of this chain as the possible maximum and disperses the starting values away from it to see if the chain will ever finish at another mode.

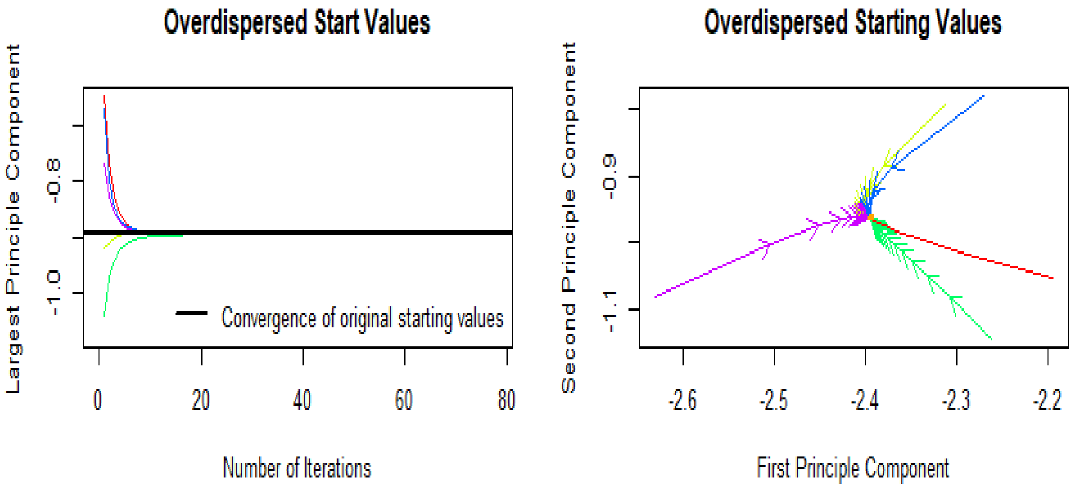

Figure 5: A plot from the overdispersion diagnostic where all EM chains are converging to the same mode, regardless of starting value.

When working with multiple incomplete variables, it is not always feasible to perform graphical checks of all imputed variables and all sets of imputations. An alternative approach is to tabulate summary statistics of the observed and imputed data. The results introduce into the Table 2. Comparing the summary statistics for observed and imputed data, the results had quite similar means and standard deviations.

Table 2: The summary statistics of the imputed data were calculated using pooled data over 5 imputations SD standard deviation, Min minimum, Max maximum

Variable	Mean	St. dev.	Min	Max	Observations
GROWTH	2.2767	3.7563	-14.2687	23.9855	486
lnGDP	10.1297	0.6959	8.2828	11.6260	486
lnPRIV	4.2523	0.5860	1.8551	5.5634	486
lnASSET	4.5894	0.0738	4.1632	4.9734	486
lnLLY	4.3286	0.6399	1.4260	6.8445	486
lnDOMCR	4.2681	0.6395	-1.6827	5.5344	486
lnSMCAP	3.4884	0.9195	-0.3047	5.5101	486
lnVALTRADED	2.0676	1.959	-3.5914	5.8753	486
lnINTEREST	0.6558	0.6637	-2.0748	2.2933	486
lnTURNOVER	3.1492	1.5696	-2.0337	6.1404	486
lnZSCORE	2.2695	0.7482	-4.0923	3.8623	486
lnNPL	1.2912	1.1230	-2.3026	3.8852	486
lnPRVOL	2.9848	0.3956	1.8467	4.1163	486
lnGINI	3.3618	0.1187	3.1091	3.5723	486
lnINCSHARE	1.5425	0.2300	1.0922	2.1187	486
lnRESDEV	0.1611	0.6498	-1.4832	1.3630	486
lnPATENT	5.3547	2.3282	-1.1087	10.1022	486
lnGOVEXP	2.9746	0.1425	2.4835	3.3299	486
lnINVEST	3.1225	0.1924	2.3240	3.7245	486
lnINFLATION	1.3520	0.9341	-2.2046	4.5146	486
lnTRADE	4.6478	0.4660	3.8159	6.0122	486
lnEDUSEC	4.6592	0.1355	4.3792	5.0995	486

4. Conclusions

This study has examined the problem of missing observations in a dataset for the 27 countries of the European Union covering the period 2000-2017, constructed to explore the relationships between financial development, economic growth, income inequality and other economic variables. The sample size is 486 observations however missing values exist in many variables. Dealing with the missing data problem, multiple imputation is the method of choice for replacing the missing values. Specifically, the statistical package Amelia II and the EMB (expectation-maximization with bootstrapping) algorithm is performed. Five fully imputed data sets are returned and the multiple imputation point estimate is the arithmetic average of the 5 sets of estimates. Additionally, Amelia II package provides several tools for diagnostics of imputation which are carried out providing accurate and valid estimates.

Acknowledgements

The paper is part of a PhD dissertation in progress, entitled “Investigating the causality between finance, economic growth and inequality using panel data” and is pursued at Department of Business Administration of University of Patras.

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: **109**).

H.F.R.I.
Hellenic Foundation for
Research & Innovation

References

[1] Honaker, J., King, G. and Blackwell, M. (2015). AMELIA II: A Program for Missing Data, R Foundation for Statistical Computing. URL: https://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

[2] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: JohnWiley & Sons.

[3] Rubin, D.B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91, 473±489.

[4] Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman & Hall.

[5] Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http:// www.stat.psu.edu/~jls/misoftwa.html.

[6] Schafer, J. L. and Graham, J. W. (2000). Missing data: Our view of the state of the art. Psychological Methods, 7, 147–177.