IMPUTATION OF MISSING VALUES IN THE FUNDAMENTAL DATA: UNLEASHING MICE FRAMEWORK

Balasubramaniam Meghanadh, Senior Quant, CRISIL GR&A, India

Lagesh Aravalath, Senior Quant, CRISIL GR&A, India

Bhupesh Joshi, Management Trainee, CRISIL GR&A, India

Raghunathan Sathiamoorthy, Manager, CRISIL GR&A, India

Dr. Manish Kumar, Director, CRISIL GR&A, India

Revolutionary developments in the field of big data analytics and machine learning algorithms have transformed the business strategies of industries such as Banking, Financial Services, Asset Management, and e-Commerce. The most common problems these firms face while utilizing data are errors, anomalies and missing values in their dataset. The major objective of the present study is to impute fundamental data that is missing in the financial statements under a multivariate framework. We use the 'Multiple Imputation by Chained Equations' (MICE) frame work to impute fundamental data by utilizing the interdependency among the variables and also complying with the accounting rules. Our proposed MICE framework utilizes the Expectation Maximization methodology in two stages with initial values based on predictive mean matching in the first stage and resolving financial constraints in the second stage to provide multiple solutions for a given line item. The MICE methodology is less time consuming relative to other techniques and allows us to implicitly enforce accounting constraints.

Key Words: multiple Imputation, MICE, fundamental data, EM algorithm; accounting and financial statement

JEL Classification: C13, C32, C51, C53, G20

Corresponding author: manish.kumar@crisil.com

Author's Profile

Balasubramaniam Meghanadh has done Bachelors in Engineering and spent six years in the engineering research and development domain. He went on to pursue Masters in Financial Engineering. In Crisil, he has been working predominantly in the areas of Risk Analytics, Quantitative Research, and Model Development for various banks and financial institutions over the past three years. He has extensive experience in Econometric Research and application of Machine Learning in finance.

Lagesh Aravalath is a PhD in Economics from the University of Hyderabad and has two years of experience in Model Validation in the areas of Consumer Analytics, Credit Risk, Economic Capital, and Market Risk. He specializes in building time-series models and has strong research acumen in the areas of economics and finance.

Bhupesh Joshi is an engineering graduate in computer science. He worked with TCS in the Analytics domain and designed web applications. He went on to pursue Masters in Economics from Madras School of Economics. He has been part of Machine Learning Innovation labs in Crisil GR&A as a management trainee and is an integral part of the team that builds prototypes for blue-sky projects.

Raghunathan Sathiamoorthy completed his MBA in Banking and Finance from United Kingdom. He has six years of experience in Model Development, Model Implementation, and Model Validation in the areas of Capital Markets Research, Multi-Manager Funds, Credit Risk, and Market Risk. He specializes in designing the framework for econometric and quantitative models.

Dr. Manish Kumar has been part of the Global Research and Analytics team of CRISIL for almost 10 years. He did his Masters and Ph.D. in Quantitative Finance in the Indian Institute of Technology Madras, India. At Crisil, he has been involved in numerous projects and assignments for various Retail and Commercial Banks, Investment Banks, and Fund Houses in the areas of Model Development, Model Validation, Regulatory Risk, CCAR and Stress Testing.

IMPUTATION OF MISSING VALUES IN THE FUNDAMENTAL DATA: UNLEASHING MICE FRAMEWORK

1. Introduction

Financial statement analysis plays an important role in investment decision making for valuation and credit analysis. The investment community uses financial statements to determine asset values, financing resources, profitability and risk embedded in the company's assets. Financial data is fundamental to asset managers and supports key investment functions such as stock selection, relative valuation, financial modelling and forecasting, portfolio construction and management, risk management, accounting and pricing of securities and successful testing and simulation of investment strategies.

Hence, the importance of accurate financial statements cannot be underestimated. Asset managers require a variety of data at regular intervals and purchase data from multiple vendors to meet their requirements. Due to the changing investment landscape, the effort required to load, cleanse, and process data has increased significantly, leading asset managers to look for more sophisticated solutions.

A key challenge faced by asset managers is securing access to reliable and clean data for analysis. Given diverse data sources, asset managers must test data for its usability. Cleaning, validating and transforming raw data to make it usable for investment research process requires considerable time and effort. Data quality is a critical issue as inaccurate data can result in costly erroneous decisions. To extract useful insights and value, the data obtained from vendors should be clean, accurate, consistent, and timely. Some typical issues with vendor data are: (1) Data inconsistency across the historical years, (2) Calculation errors, (3) Data duplication, (4) Vendor policy or definition inconsistencies for key data items and (5) Missing or incomplete data.

Problems with data also arise because firms frequently source data from various third-party vendors and each of these vendors has its own data definitions and data structure. Firms also maintain their own databases which again have different data definitions and data structures. All data, both real-time and historical, contains errors and issues leading to unreliable and wrong investment decisions, inefficiencies, increased operating costs, missed opportunities and reputational damage.

Andrea et al. (2012) [1] analyzed balance sheets in the banking sector and noticed that some data values are either missing or incorrect. They suggested using a forward search algorithm to identify the incorrect observations. The forward search algorithm progressively adds each data point into the dataset and searches for deviations in the point estimates to identify anomalies. We can treat these anomalies as a missing data problem as incorrect data cannot be used for any analysis and requires the same treatment as missing values.

Rubin (1987) [2] notes that ignoring missing values for analysis will introduce bias if a systemic difference between respondents and non-respondents (observed and missing data) exists in the datasets. Further, missing values reduce the size of the data and ignores the uncertainty in the missing data causing biased estimates. Kofman & Sharpe (January 2000) [3] notices that in key research papers (Journal of Banking and Finance, the Journal of Finance, the Journal of Financial Economics, the Journal of Financial and Quantitative Analysis and the Review of Financial Studies), the practitioners have altogether ignored and not reported the missing value problem by noting that only 175 of the 1,057 articles explicitly recognized the treatment of missing data. Even in 137 of those 175 articles, the authors have used the list wise deletion method which removes observations containing missing data as their primary data treatment.

The impact of not handling missing data effectively can have serious consequences on quantitative research, leading to: (1) biased estimates of parameters, (2) loss of information, (3) a decrease in statistical power, (4) an increase in standard errors, and (5) weak generalizability of findings. Hence, imputations of missing values are always considered a better choice than removing them. For example, Joos et al. (1998b) [4] suggested that the outliers and missing values should be corrected in the dataset before performing the model building exercise. Despite the availability of various imputation methods, suitability of the chosen method is based on the specifics of the business problem.

Rubin [2] is considered a pioneer for identifying a technique called multiple imputation to address the missing value problem. The importance of this approach is that imputation is performed several times and the practitioners can perform modeling and analysis on multiple datasets to present their final conclusions. Fogarty (2006) [5] used the multiple imputation approach by treating the reject inference as a missing data problem to enhance the information inferred from the rejected applications over the traditional reject inference approaches while developing credit scorecards. Galler & Kehral (2009) [6] performed multiple imputation to assess the probability of default of a given company using financial ratios as independent variables. Bouhlila & Sellaouti (2013) [7] performed missing value imputation for the student background data file for Tunisia using the multiple imputation approach. Chen (2013) [8] suggested the multiple imputation method to defuse the lack of variability in the data field substitutions for banks to treat the missing value problems and stresses that banks should not consider missing data an insurmountable problem and address it.

A major objective of this study is to develop a framework to handle fundamental data that is missing and has errors, inconsistencies and anomalies. The important aspect of our study is to utilize existing research on the MICE framework that incorporates the interdependency among variables and also enhances the framework to impute missing data that also complies with accounting rules (i.e., principle to fundamental data). For example, aggregation of line items within the framework of asset, liabilities, and owners-equity of the balance sheet should be matched with the accounting principle (Assets - Liabilities = Owner's Equity). The line items that are on the balance sheet could possibly have a relationship with the cash-flow statement and the income statement, and the relationship should also be handled as a single-fold problem with the MICE framework. For example, operating cash flow, investing cash flow and financing cash flow should be tallied with the net-cash difference in the balance sheet of the previous year. Given the complexity of accounting constraints, they should be implicitly handled within the framework, the construction of rule/accounting constraints is based on the interaction among line items of the financial statement.

The current literature review suggests that traditional imputation techniques such as mean/median imputation, list wise deletion, and omission of variables cannot accommodate the accounting constraints and interdependency among line items in the financial statements as they show a clear time trend and mean/median imputation cannot provide appealing results. Thus, in our study, we focus on imputing the values subjected to financial constraints post identification of the anomalies in the balance sheets, income statement and cash flow statement using the multiple imputation approach. Kofman & Sharpe (January 2000) [3] compared advanced techniques such as Expectation Maximization (EM) algorithm and Imputation Posterior method (IP) within the Multiple Imputation by Chained Equations (MICE) framework. The disadvantage of IP over EM methodology is that the former is computation intensive in achieving convergence (Kofman & Sharpe, January 2000) [3]. Hence, our approach is based on the Expectation Maximization Algorithm suggested by Dempster, Laird, & Rubin (1978) [9] which iteratively performs missing value imputation and estimation of parameters of the distribution till the desired converge is obtained for the distribution (maximum-posterior).

The paper explains the MICE (multiple imputed chained equation) approach using EM algorithm to solve the missing value problem in financial data. Chained equations refer to the fact that the MICE algorithm can be easily implemented as a concatenation of univariate procedures to fill missing data. The beauty of the MICE approach is that it handles the links between different financial items seamlessly. Effectively, with only a couple of line items such as Total Assets and Net Sales, the imputation can be performed for line items at all levels till the most granular item in the form of a tree structure using this approach. The EM algorithm is computationally very efficient and performs the imputation for about 177 line items in less than a minute for a given company. We have used a performance metric similar to that used by Galler & Kehral (2009) [6] to gauge the beta coefficients for each of the line items. Additionally, we have used line charts and ratio comparison between the previous and current year to assess the performance of imputed values.

The remaining sections of this paper are organized as follows: Section 2 provides a review of the relevant literature. Section 3 discusses the data. Section 4 details the methodology. Section 5 reports results, including the algorithm, assumptions, limitations and the challenges faced while using this technique. Section 6 conclude the study by providing a brief discussion on how our research could be useful for practitioners.

2. Literature Review

Kofman & Sharpe (January 2000) [3] notice the following missing value imputation techniques in financial journals used by the practitioners such as list-wise deletion method, omitted variable method, ad-hoc methods such as mean imputation, randomly assigning zeroes and ones to categorical data, etc. The authors pointed out the disadvantages of each approach. For example, the list-wise deletion method may introduce a systematic bias to estimates when the observations with missing data appear to have characteristics more attuned to a particular outcome of the dependent variable. In the simultaneous equation studies, the missing data exclusion method is known as the pairwise deletion method. The method has limitations such as inconsistency in the covariance matrix, biased estimates and small standard errors. Finally, they compared the results of list-wise deletion methods with the EM methodology and assessed that the EM methodology performed better and the list-wise deletion methodology was vulnerable to significant negative bias.

There are many imputation techniques that are commonly used to deal with the missing values that fall under the missing at random (MAR) category. These methods can be classified as single imputation techniques (SIT) and multiple imputation techniques (MIT). In SIT, the missing values are replaced by some type of "predicted values" from the available cases/information. In SIT, the missing values are imputed once; hence, it is called single imputation techniques.

Rubin (1978, 1987 and 1996) [9], [10] and [11] proposed the multiple imputation technique (MIT) which overcame the limitations of SIT. Multiple imputation involves three phases: imputation, analyses and pooling. In the imputation phase, the process is iteratively designed to arrive at multiple optimal values for a given variable. In the analyses phase, the desired analyses are performed on each dataset using standard complete-data methods. In the pooling phase, all the multiple results are consolidated by calculating the mean, variance, and confidence interval of the variable of interest. One major advantage of MIT is its flexibility and ability to accommodate various scenarios. Also, MIT can be used where the data is missing completely, missing at random, and missing not at random.

MICE is a multiple-imputation technique proposed by White et al (2011) [12] and based on the multiple-imputation technique suggested by Rubin D. (1987) [10]. It is built based on the assumption that the missing data is MAR; hence, the probability that a value is missing depends only on observed values and not on unobserved ones. Fundamental data follows clear rules and hence can be considered MAR. Also, the unobserved variable is a direct function of an observed variable. MICE can incorporate datasets with thousands of observations and hundreds of variables

(He et al (2009) [13]; Stuart et al (2009) [14]). In the MICE method, each variable is modeled based on the assumptions and characteristics of the underlying distribution. The fundamental data is continuous; hence, the MICE methodology can be applied to this problem.

Little & Rubin (2002) [15] and De Waal (2011) [16] detail the following single value imputations such as mean imputation, cold deck imputation, hot deck imputation, and regression imputation in practice. The common problem in single imputation is to replace an unknown missing value by a single value and then treat it as if it is a true value Rubin D. (1987) [10]. As a result, single imputation ignores uncertainty and underestimates variance. Multiple imputation overcomes this problem, by taking into account both within-imputation uncertainty and between-imputation uncertainty. Fogarty(2006) [5] explains that multiple imputation retains the advantages of single imputation and rectifies its major disadvantages by replacing each missing value with a vector composed of $M \geq 2$ possible values. The vectors of imputations create a $n_{mis} \times M$ matrix of multiple imputations, where n_{mis} is the total number of missing values. Each column of this matrix is used to create a completed dataset, and so multiple imputations lead to M completed datasets, each of which can be analyzed using statistical or data mining techniques that are appropriate when there are no missing values. For example, using standard complete-data methods, an analyst can obtain *M* estimates and their variance-covariance matrices or *p* values, which can then be combined to form a single inference under the model used to impute the missing values ((Li, Raghunathan and Rubin 1991 [17]; Rubin and Schenker, 1991 [18]). Schenker et al., (1993) [19]) have reported that multiple imputation has been developed and can be justified most easily from the Bayesian perspective. In multiple imputation, two methods are used by the analysts which are joint modeling (JM) and fully conditional specification (FCS), otherwise commonly known as MICE (Multiple-Imputation by Chained Equation) Buuren (2007) [20] and Raghunathan, (2001) [21]. FCS allows imputation on a variable-by-variable basis and hence is preferred over the JM approach.

Buuren, et al., (2006) [22] explains several advantages of FCS over JM. The major advantage of FCS is its increased flexibility in model building. It is easy to incorporate constraints on the imputed values, work with different transformations of the same variable, account for skip patterns, rounding, and so on. Specifically, unlike FCS which requires only less than 10 iterations, many Markov Chain Monte Carlo methods (JM method) often require thousands of iterations.

Our MICE approach is based on the EM algorithm suggested by Dempster et al., (1978) [9]. Kalb et al. (1995) [23] also uses the EM algorithm to perform imputation on missing values in a single equation for actuarial applications. Malhotra, (1987) [24] used this algorithm in the marketing research area for probit and tobit estimation problems. The advantages of EM algorithm are that convergence happens without any assumptions on the derivative functions or the starting values and it occurs for even small sample sizes when the EM parameters are single values (maximum posterior) instead of a complete distribution. In the multiple imputation framework, which is a variable by variable imputation for all the variables in a specific order, m possible alternative datasets are presented and the final estimates are pooled together; hence, this may not be a bigger issue. Fogarty (2006) [5] used the multiple imputation technique to build scorecard models by treating the reject inference problem as a missing value problem for declined applications. This is because Henley (1997) [25] describes the importance of robust technique for reject inference modeling as using the accepted applications alone will introduce a systemic error in scorecard applications. Takahashi & Ito (September 2012) [26] used Expectation Maximization with Bootstrapping algorithm to perform missing value imputation in economic census data where the dependent variable is the turnover in different sectors and the number of workers is the independent variable. They confirmed that multiple imputation is closer to the true values than single imputation. Bouhlila & Sellaouti (2013) [7] obtained smaller standard errors and narrower confidence intervals along with the advantage of using the entire data set while performing multiple imputation.

Tanner et al. (1996) [27] used a variation of EM method called EM-sampling, where the mean and variance are estimated in step 1. In step 2, the normal distribution is constructed based on the estimates derived from step 1. In step 3, the point estimate is derived from a normal distribution and used to provide the initial value. The process is repeated iteratively to provide the final distribution of imputed values. Gary et. al. (1998) [28] shows that this works well for a normal distribution, but with highly skewed (non-normal) categorical data, it can produce incorrect standard errors. This disadvantage can be corrected in the EM-importance sampling methodology using an importance ratio which takes into account the likelihood of the estimate while choosing the samples based on the observed values.

Kofman & Sharpe (January 2000) [3] have compared two advanced machine learning approaches for multiple imputation which are EM-importance sampling (EM-*is*) and Imputation-Posterior (IP) methods. The author prefers the EM-*is* over IP as the latter is computationally intensive.

Another choice that a practitioner should consider while performing missing value imputation using the EM algorithm is to identify the correct model specification. Gary, et al, (1998) [28] suggests that the missing value imputation model need not be the "*analysis*" model as the risk of over specification is not a concern. However, we have used the analysis model itself as the alternative technique for missing value imputation as suggested by Schafer(1997) [29] and Raghunathan(2001) [21].

3. Data

Most asset management strategies are built using fundamental analysis of statements based on earnings, sales, debt, cash flows and related metrics such as profitability ratios, liquidity ratios, debt ratios, and efficiency ratios. They are also used to estimate a company's liquidity and default risks. Asset managers also use quantitative screeners to reduce a large investible universe into a smaller group of stocks, and then apply fundamental analysis to shortlist potential investment opportunities. Historically, asset managers have typically relied on internally developed technology solutions to manage and obtain all the required data and information but they currently use third-party solutions to meet their data requirements.

Fundamental data enjoys several unique features and is different from traditional statistical data which is typically based on a marketing survey or a field survey. As the research study focuses on fundamental data, the major challenges that the imputation should accommodate are interdependencies between line items and adherence to accounting constraints. Also, the study used the fundamental data sample of US-based companies in the healthcare industry sourced from a third-party vendor. The data comprises fundamental financial data of about 1,000 line items containing various information about the companies, including the balance sheet, income statement, financial ratios, equity information and market related information. The data is collected in both quarterly and annual frequency spanning a period of six years starting from 2010 through 2016. However, for 2016, only a few financial ratios are available and none of the top-level items such as Net sales and Total assets are available. However, this cannot be treated as a forecasting issue as the top-level items can be derived from the financial ratios directly or indirectly (which should be taken care within the algorithm) and these ratios are good enough to impute the rest of the line items using the framework developed in this study.

The present study imputes values for a total of 177 line items in the fundamental data across the categories viz. Balance sheet, Income statement, Cashflow and financial ratios. Of the 177 line items, 73 items are from the balance sheet, 36 items are from the income statement, 40 items are from cashflow and 28 items are financial ratios. The values in the fundamental dataset for 2016 are not available due to various reasons. Also, preprocessing of the dataset before modelling is recommended to capture errors, inconsistencies and anomalies. These anomalies flagged earlier are removed from the data set as they may not be useful for further analysis.

Imputation of missing values was based on the historical pattern in line items. There were several challenges in imputing fundamental data, as the data should satisfy all the accounting principles or constraints. Each and every line item can have multiple constraints. However, if we provide all these constraints along with rounding-off errors, a single optimal solution may not be feasible. However, all the line items should respect the accounting rules, otherwise the imputed values will not be useful. In our algorithm, we have used only one constraint for each line item and still manage to adhere to all the accounting rules.

The constraints for each line item are not static; hence, they will be chosen based on the non-missing values in the data set. Further, we cannot pre-fix the constraints, as the non-missing items could be different for a different company. The constraints should also be able to perform simple imputation if all the concerned line items are directly available. i.e., the constraints should be specified in such a way that if the value for line-items that are contributing to a particular line item are directly available, then the missing values for that line-item must be directly solved. For instance, if a line item ‘a’ can be specified as the difference of the two other items i.e., ‘b’ and ‘c’ and the values for ‘b’ and ‘c’ are directly available, then the missing value for ‘a’ can be solved directly by providing the constraint a=b-c.

Also, only the top-level line item was available (Total Assets) in the balance sheet; hence, we adopted a top-down approach to derive the relationship to impute the lower-level items. Further, the cash flow line items should be derived from the balance sheet and profit and loss statements in addition to the rules for the cash flow statements. Also, some line items are fundamental ratios that need to be imputed directly. Moreover, all the imputed values should follow all the fundamental ratios for which values are non-missing. Further, assigning appropriate units to each line item is a challenge, as different line items have different units such as $million, ratios or percentages. We were able to tackle this issue by properly specifying the unit for each line item in the dataset.

Hence, incorporating the constraints within the Machine Learning algorithm is a challenging task. The MICE technique is capable of handling these constraints effectively Buuren et al. (2006) [22]. The challenge in this study is that each line item must have a specific constraint. However, all the financial items have constraints and hence using the MICE approach is still a challenging task. This study develops a framework to solve the missing value imputation problem in fundamental data.

In MICE, imputation is performed in a chained manner. We have fixed the order sequence of imputation based on specific levels. For this purpose, we have assigned a 'level' for each financial line item. In our study, all the line items had taken levels from level 0 to level 6. Based on these levels, a specific ordering sequence is provided in MICE. As we are following a top-down approach, the ordering sequence followed is from level 0 to level 6. This 'level' variable will take care of the order of solving the constraints according to their importance. The first level items will be solved in the first stage, then the next level is addressed and so on. Level 0 either means the value is directly available or it can be directly computed from all the concerned variables. Level 1 means there is one top-level item that is available and there are multiple lower level items; hence, multiple imputation has to be performed. For example, if there is an equation A=B+C+D, then if A is available and B, C, D have to be imputed, then A is called a Level 0 variable and B, C and D are called Level 1 variables. As there are multiple line items in Level 1, we have provided another classification called 'importance' which is assigned based on the historical values for the line items B, C and D. For example, if B>C>D, then the importance for B, C and D are 1, 2 and 3, respectively.

Rubin (1976) [9] classified missing data into three categories, namely missing completely at random (MCAR), missing at random (MAR) and Missing not at random (MNAR). In MCAR, missing values are randomly distributed across all observations. A missing value is MAR if the probability that a value for a certain variable is missing is related to observed values on the other variables or it can be explained by other variables. In this case, the missing values are not randomly distributed across observations but are distributed within one or more sub-populations. This is one of the very common types of missing values observed. This type of missing values can be modeled using observed variables. A missing value is MNAR when there is a probability that a missing value is unrelated to the values of observed variables. (Schafer, 1999, p.8) [30]) mentions that the assumption of missing at random (MAR) is not required in multiple imputation. The financial data is strictly governed by the accounting rules and the top-most level item is not imputed in this approach. Hence, MAR is a reasonable assumption for fundamental financial data.

4. Methodology

The multiple imputation technique is very intuitive and statistically appealing. The importance of this approach is that it performs the imputation with a random draw from a given distribution multiple number of times. Hence, we will have multiple datasets at the end of the imputation that shows the multiple optimal solutions can be outputted. The practitioners can perform the modeling, analysis on the multiple datasets instead of a single data set using a process called ‘pooling’ to present their final conclusions.

4.1.Missing value imputation techniques

Let the hypothetically-complete data Y be a partially-observed random sample from the p- variate multivariate distribution $\mathrm{P}(\mathrm{Y}/\theta_1)$. We assume that the multivariate distribution of Y is completely specified by θ, a vector of unknown parameters. The problem is to get the multivariate distribution of θ, either explicitly or implicitly. The MICE algorithm obtains the posterior distribution of θ by sampling iteratively from conditional distributions of the form

$$\mathrm{P}(\mathrm{Y}_1/\mathrm{Y}_{-1}, \theta_1)$$

$$\mathrm{P}(\mathrm{Y}_p/\mathrm{Y}_{-P}, \theta_p)$$

The parameters θ_1,…. θ_P are specific to the respective conditional densities and are not necessarily the product of a factorization of the `true' joint distribution $\mathrm{P}(\mathrm{Y}/\theta_1)$. Starting from a simple draw from observed marginal distributions, the *t*th iteration of chained equations is a Gibbs sampler that successively draws

$$\theta_1^{*(t)} \sim P(\theta_1/y_1^{obs}, y_2^{(t-1)}, \dots..y_p^{(t-1)}$$

$$y_1^{*(t)} \sim P(y_1/y_1^{obs}, y_2^{(t-1)}, \dots..y_p^{(t-1)}, \theta_1^{*(t)}$$

$$\theta_P^{*(t)} \sim P(\theta_P/y_P^{obs}, y_2^{(t)}, \dots..y_p^{(t)}, \theta_P^{*(t)}$$

$$y_P^{*(t)} \sim P(y_P/y_P^{obs}, y_1^{(t)}, \dots..y_p^{(t)}, \theta_P^{*(t)}$$

where $y_1^t = y_j^{obs}, y_j^{*(t)}$ is the *j*th imputed variable at iteration *t*. We observe that previous imputations $y_j^{*(t-1)}$ only enter $y_j^{*(t)}$ through its relation with other variables, and not directly. Convergence can therefore be quite fast, unlike many other Markov Chain Monte Carlo (MCMC) methods. EM Algorithm is used to find θ that maximizes $g(x/\theta)$ given an observed y. Let $f(x/\theta)$ is a family of sampling densities, and

$$g\left(\frac{x}{\theta}\right) = \int_{F^{-1}(y)} f\left(\frac{x}{\theta}\right) dx$$

The EM algorithm aims to find a θ that maximizes $g(x/\theta)$ given an observed y, while making essential use of $f(x/\theta)$. Each iteration includes two steps: (1) the expectation step (E-step) which uses current estimate of the parameter to find (expectation of) complete data. (2) The maximization step (M-step) which uses the updated data from the E-step to find a maximum likelihood estimate of the parameter. It stops the algorithm when change of estimated parameter reaches a preset threshold.

4.2.Imputation of missing values in the Fundamental Data

Functional form of a line item

In our approach, we have used the predecessor as a predictor variable (either raw predecessor, or difference in the predecessor or growth of the predecessor). For few line items, these relationships may not hold. This leads to the process would not provide optimal solution to solve for constraints. This problem is because of the error in the specification. As this is rare scenario, this is considered as limitation for the model and the severity is justified. For example, a line item unexpected gains/loss may not follow a specific distribution nor can be predicted with its predecessor. Unless we have a unified approach, those line items limit the model to arrive at optimal solution. Therefore, we have duplicated or shadowed all the line items leading the model to pass through two-fold stages. In the first stage, one of the variables is predicted based on the relationship with the predecessor. In the second stage, the duplicated line item is imputed based on the prediction from the first stage simultaneously solving for accounting constraints.

For e.g.: Sales = Gross Income + Depreciation, Depletion and Amortization + Cost of Goods Sold

Then, Gross Income should solve the constraint as mentioned below

Sales - Depreciation – Cost of goods sold

The mathematical process of the two stage approach for the above example is given below

$$\text{Stage 1: } Gross\ Income = \beta_0 + \beta_1 \times Sales$$

$$\text{Stage 2: } Duplicated\ (Gross\ Income) = \beta_0 + \beta_1 \times Gross\ Income$$

If the Gross income has to follow the relationship with its predecessor, it might be possible that it may not solve the constraint for any specific reason. In this case, solving for the constraint is more important than following the distribution, as it is merely an estimate. Moreover, the β_1 in the second stage is expected to be closer to 1. Hence, the deviations in the β_1 in the second stage from its desired value is an indication that the distribution of the imputed value which is solving for the constraint is deviating from the desired distribution. By verifying β_1, we can verify the mis specification and improve the relationship of the line items by selecting the right predecessor. We have used the skip-level predecessor as a predictor variable in few cases, where the correlation with the skip-level variable is higher than the correlation with the immediate predecessor.

Tree Structure

The order in which the variables are imputed is very important while performing a variable-by-variable imputation. If the order of imputation of variables is not considered, then it may introduce circular constraints that cannot be solved. To avoid this incompatibility, we have introduced the ‘tree structure’ approach to identify the order of imputation. This idea stems from the fact that for financial statements like balance sheet, income statement and cash flow statement we have a top-most level item categorized namely Total Assets, Net Sales, and Net cash respectively. Using Asset Turnover ratio that links both the total assets and net sales, we have prepared a single tree structure for all the line items based on its predecessor. For example, we consider Asset turnover ratio as a Level 0 item. Both Total Assets and Net Sales are level 1 item. The components that add up to the total assets and net sales are level 2 variables and so on. With this mapping we can map each and every item in the Balance sheet, Income Statement and Cash flow statement to a specific level.

Constraints

The relationships for a specific line item in the tree structure are explained in this section. Let us assume that we are interested in the relationships for a line item called gross income. Gross income is derived from the sales. Hence, sales is the predecessor of Gross income. As per the accounting rules, net sales is the aggregation of Gross Income, Depreciation, Depletion & Amortization and Cost of Goods sold. In the top-down approach, we are not providing this constraint to the Net Sales. However, each of the Gross Income, Depreciation, Depletion and Amortization (DDA) and Cost of Goods sold (COGS) will take this constraint by rewriting the above equation for each of them.

A major challenge in identifying a methodology that can be used to impute the missing values in the ‘fundamental data’ is it should be able to simultaneously maximize the maximum likelihood function and solve for the given constraints. Accounting constraints are more challenging for a single line items having multiple relationships.

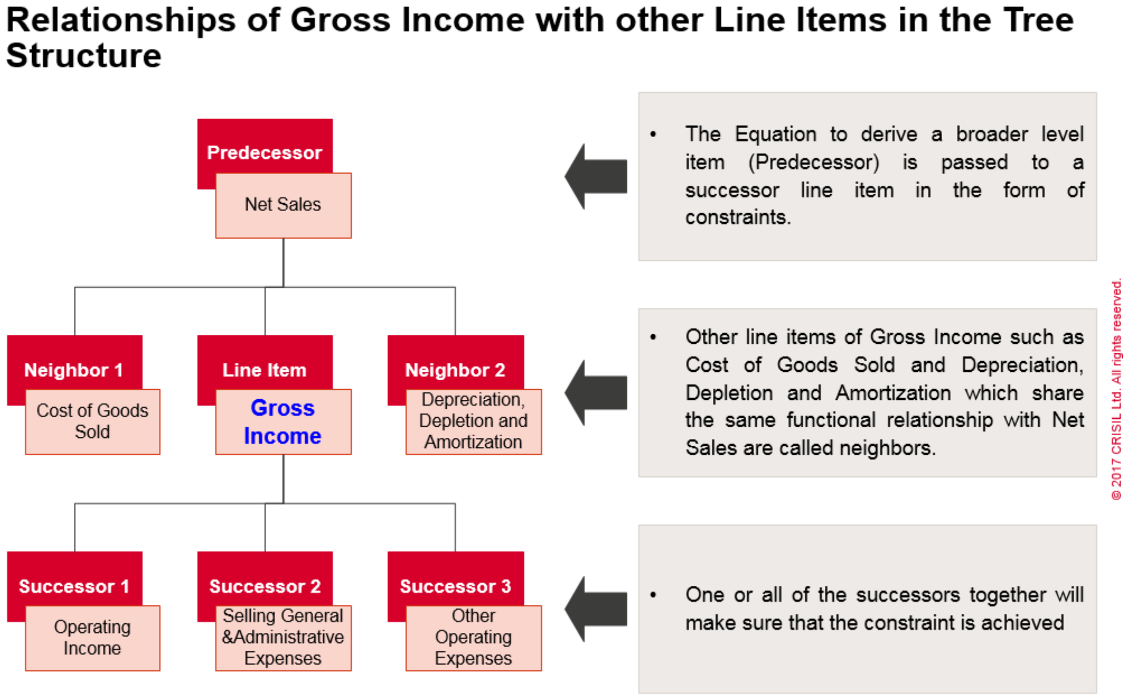
The algorithm should respect the predecessor and successor’s relationship effectively. Any line item has three relationships:

(1) The predecessor, which is typically a top-level item. However, if few ratios are directly available then the predecessor will be changed to that specific ratio.

(2) The successors, which are typically lower-level items or in other words, which require the imputation of that specific line item.

(3) The neighbors, which share the same constraint linked by a common equation. For example, total debt is equal to the sum of short-term debt and the long-term debt. Both short-term and

long-term debts are neighbors to each other as they share the same relationship. With this approach, we are able to preserve the relationships for all the line items.



Relationships of Gross Income with other Line Items in the Tree Structure

- The Equation to derive a broader level item (Predecessor) is passed to a successor line item in the form of constraints.
- Other line items of Gross Income such as Cost of Goods Sold and Depreciation, Depletion and Amortization which share the same functional relationship with Net Sales are called neighbors.
- One or all of the successors together will make sure that the constraint is achieved

Figure 1: *An example of the relationship of a line item (Gross Income) with the other line items in the tree.*

The algorithm should also keep the tree structure intact even if few line items can be directly derived from the financial ratios. Our approach is able to efficiently take care of this situation. Our algorithm works as follows:

(1) With the initial tree structure, the line item Net Income – Bottom Line takes a level of 5 derived from the Level 0 predecessor which is Net Sales or Revenues

(2) However, for most of the companies, Profit to Sales Margin is available. With the virtue of this relationship, Net Income – Bottom Line can take level 1 as it can be directly derived once Net Sales is imputed

(3) Since the neighbor of Net Income – Bottom Line, Extraordinary Item Gains/Loss from Sale of Assets share the same constraint as that of Net Income, the tree structure will still be preserved

Our approach is very intuitive and is solving the constraints efficiently. For practical purposes, the imputed values may not pose any serious issue. Further, all the imputed values will be flagged so that all the data users are aware of the exact location where the data is imputed.

4.3.Advantages of MICE approach

The MICE approach has several advantages:

(1) It is built on sound Expectation Maximization algorithm which is an unsupervised learning method. Only assumption in this algorithm is that the functional form represents the precise relationship between the independent and dependent variables. As the financial ratios are mostly stable across time, the regressor coefficients will be stable and the assumption will be valid for fundamental data. Few variables follow the growth pattern instead of ratios. For example, we use growth in sales, operating cash flow, current assets, and etc.

(2) The algorithm automatically identifies the parameter which represents the best fit. Further, once the predecessors are imputed, the Expectation Maximization algorithm will re-adjust the coefficients and re-estimate the lower-level item. After the parameters are chosen, the EM algorithm will assign specific values for the dependent variable. Based on the assigned values, the parameters are recalibrated until the desired convergence process is achieved. Hence, the possibility of feedback mechanism is embedded within the algorithm itself.

(3) Providing financial constraints is possible within the MICE package. R is an open-source software with contributions from several practitioners. It is extensively documented and several reviews are available on the packages from the contributors. R has a MICE package which is built on the above-mentioned methodology and the users can leverage and build upon the software to meet their individual needs. The MICE package in R has the capability to take the constraints as inputs and solve them within the imputation process itself.

4.4.Assumptions

The assumptions of the methodology are given below

(1) The financial ratios between the dependent and independent variables is assumed to be constant during imputation process. We assume that there are no corporate actions like buy-back of shares, M&A, etc. that could impact the financial ratios.

(2) The dependent variable can be imputed with one variable. This assumption is valid as we have very few data points and there is only one predecessor for each line item

(3) For few lower-level items, even if the values are slightly away from the distribution it is still not a big concern as these values are following the financial constraints. For example, if most of the values are zero (except one or two), then even if the imputed values are not zero, it does not pose a serious threat

(4) The predictors provided, based on the predecessor relationships, will hold for all the companies in all the sectors

(5) The input data is of high quality

4.5.Algorithm

The algorithm takes the following steps:

(1) The data for a specific company ID is retrieved from the base file and all the non-missing values are marked as level 0

(2) A constraint sheet is prepared which has the following information: Field ID, Field Name, Level, Predecessor and a constraint equation. The levels, predecessors and constraint equations are initially designed based on the world scope balancing model. For example, sales and total assets are level 0 items and so on. We note that the level, predecessor and constraint equations are dynamically selected within the algorithm.

(3) Data pre-processing steps are carried out to obtain data for a specific company in a specific sector.

(4) For initial imputation, we use only one variable which is the predecessor. The assumption is that all the lower level items have a specific ratio relationship with its predecessor. This is true in most of the cases. However, cash flow items are also derived from income statement and balance sheet which depends on previous year relationships. For example, an increase or decrease in cash is calculated from the difference of cash in the balance sheet between the two years. For such cases, we assign the predictor variable as difference of the cash itself. For few other items, we provide growth of predecessor as the predictor variable.

(5) For few cases, the relationship between the predecessor and successor does not hold true. As MICE will impute the values based on the distribution, the solutions in few rare cases are not

feasible. To avoid the convergence criteria, the algorithm has been modified slightly so that we duplicate each and every line item-one item will perform the imputation based on the predictor and the other one will be solved for the constraint by slightly moving away from the original distribution. By this way we are able to impute all the line items in the balance sheet, income statement and cash flow statement and by incorporating few important ratios would derive the broad level items.

(6) The levels, predecessors and constraint equations are updated for each company based on the data that is available for the specific company. We intend to promote each and every line item based on the non-missing data. Accordingly, the levels, predecessors and equations are updated simultaneously without breaking the already existing relationships.

(7) Using EM algorithm, the MICE package performs the imputation from a univariate distribution for every company (as we use one variable as a predecessor). The functional form is a one-variable linear regression with the predecessor as a predictor variable and simultaneously solving that value for the given constraints. MICE package in R has an inbuilt option of choosing a value from within the distribution itself. The option is known as predictive mean matching. This is same as linear regression, except that instead of imputing the predicted value, the imputation will be performed based on a value within the distribution which is very close to the predicted value. In essence, the imputed value will be from the distribution itself. The option of choosing predictive mean matching (pmm) is intuitive. As it is a primary requirement that the imputed value should follow distribution of the other non-missing values, using 'pmm' imputation is advantageous. For cases where the data clearly displays a time trend, by modeling the difference of a dependent variable (such as AR (1) process) using difference of independent variable, we can achieve the desired performance. The independent variable is diff(x) and the dependent variable is diff(y).

(8) We verify the plot of the line items in an interactive chart and check the reasonableness of the imputed value.

4.6.Challenges

Given the assumption that the missing values are at random, the values could be missed at different levels for different stocks/firms. Also, few top-level items could be available. The dependency on the financial ratios, and dynamic nature of predecessor-successor relationship demands the framework to incorporate these changes across stocks in different sectors. For E.g.: The line item that is available for a given stock may not be available for another stocks, hence the structure of the tree has to be dynamic in nature. The complexity of accounting constraints, and interdependency among variables have to be constructed with the subject matter experts in financial statement analysis. The financial constraints are constructed based on ratios and % of line items, thus leading to intricacy in handling the relationship between cash flow and balance sheet. As the line items could be missed at random, the top-level items could not be reported for few companies. For E.g.: If cash is the only component of current asset, they might end up reporting Cash as top-level item.

4.7.Limitations

For few companies, the relationship of the financial ratios between the dependent and independent variables is volatile and cannot be used as a component for modeling. When the line items are imputed based on the financial ratios, and the imputation of this could affect accuracy. The dependency on the financial ratios could be solved by the availability of top level items. Both the top-level items and financial rations missing simultaneously is rare. Hence, the severity of this limitation is deemed low. If the model is intended to forecast, the severity could be high. Despite the most granular items at lower levels have multiple predecessors, the nature of the fundamental data enables the framework to prefer fundamental ratios for the imputation. Also, the imputation is based on growth rate or AR (1) as a functional form. The few line items that follows the trend could be distorted from the distribution, and the severity of this limitation is low. Any unusual trend could be observed by the visual charts in time-series plots. Despite a weak relationship between line items and its predecessors, MICE enables predict mean matching to impute the actual value from the distribution. Hence, severity of this limitation is low. The identification of data anomalies should be a separate exercise before processing the data within the MICE framework. The data anomalies can be flagged using rigorous data analytics by comparing with other third-party data, outlier analysis based on distribution, and many visual presentation of data set. These anomalies can be considered as missing, and processed similarly within MICE framework.

5. Result and Discussion

We have performed imputation for 177 line items using the MICE framework and adhering to all accounting rules. Overall, we have 73 constraints for various line items that are solved using this approach. We have used line charts and compare the financial ratios of the previous year and the current year to assess the reasonableness of the imputed values for the year 2016. The results indicate that the approach can handle such a large problem very efficiently.

Table 1: Percentage of constrains Solved using MICE.

Total Number of Line Items	**Total Constraints**	**Number of constraints solved**	**% of constraints solved**
177	73	73	100%

Further, to validate the MICE approach, we have performed the out-of-time testing. Wherever we have missing values for 2016, we have removed the corresponding values for 2015. We have also removed 2016 data as most of values are only imputed values. We have used the same setup that has been used to resolve the missing values for 2015 to impute values for 2016. These values are compared with the actual values and we have calculated the error and percentage error. We have grouped the differences into five categories and provided the results in the table below. The results are encouraging given that we have adopted a relatively simple approach instead of developing a model for every line item.

Table 2: Out of Time Test- Performance Metric: (Left) Percentage Error; (Right) Error

Percentage Error Table	
Range	Line Items
-80% to -50%	8
-50% to -10%	21
-10% to 10%	127
10% to 50%	14
50% to 80%	7

Error Table	
Range*	Line Items
Below -500 million	2
-500 to -50 million	18
-50 to 50 million	140
50 to 500 million	16
Above 500 million	1

* Note: The total sales of the company in the year 2015 is $9,800 million and total assets is $21,500 million

Further, we have compared the difference in the model estimates in out-of-sample testing. We have compared the results of one of the many (10) imputed datasets with the results of the models based on the actual data for the year 2016, following Galler & Kehral (2009) [6]. These results are provided below:

Table 3: Difference in Beta of slope between actual and the imputed data sets

Difference in Beta estimates of slopes between actual and imputed data sets	Line Items
0	8
0.4	133
0.8	21
1.5	10
>1.5	4

Table 4: Difference in Beta of intercept between actual and the imputed data sets

Difference in Beta estimates of intercepts between actual and imputed data sets	Line Items
0	9
10	29
100	132
1000	7

Further, deviations of point estimates from its desired values (desired value is Beta=1) in the second stage model for duplicated variables are also presented below.

Table 5: Difference in Beta estimates in the Second stage between imputed values and the expected values

Difference in Beta estimates in the Second stage between imputed values and the expected values	Line Items
0	27*
0 to 0.3	0
0.31 to 0.6	0
0.61 to 0.8	9
0.81 to 0.95	16

0.96 to 1.05	92
1.06 to 1.2	27
1.21 to 1.4	6

* Note: Difference in Betas is Zero as the actual values are zero

Further, we have presented the line charts of the few imputed line items for the year 2016 below.

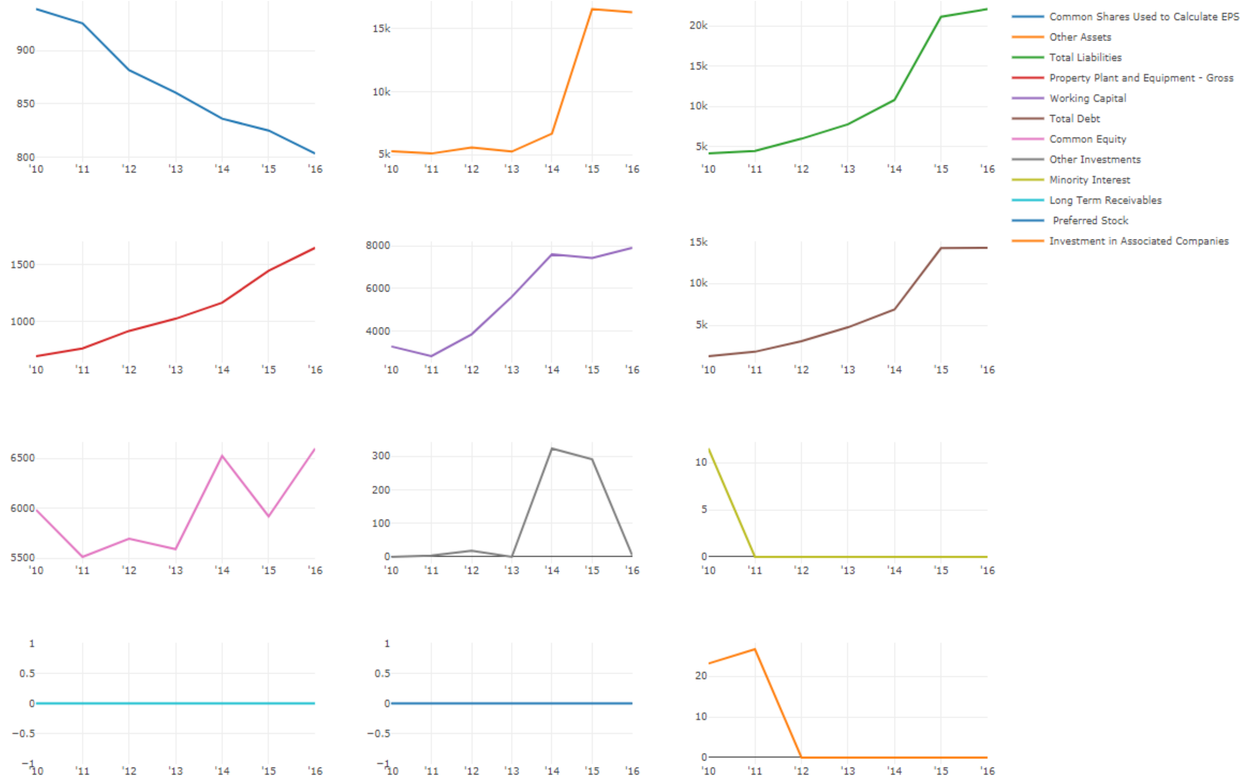


Figure 2: *Sample of imputed Line Items for 2016 (2010-15 are Historical values)*

6. Conclusion

In this study, we have utilized Fully Conditional Specification (FCS) to solve the missing value problem in the financial dataset using a technique prescribed in standard literature (especially Rubin, Schaefer and Raghunathan) and performed imputation for 177 line items (out of which 73 is missing). The results are very appealing given that the number of observations required to train the model is minimal. Also, the process of identifying the functional relationship between different line items requires functional knowledge in handling financial statements. Our research studies the application of MICE suggested in the literature by imposing accounting constraints and also capturing interdependency of line items within the components of financial statements. The functional relationships can be improved by leveraging the experience of asset managers and the approach can be customized for various line items to improve the efficacy of the imputation process.

We have developed a framework within MICE such that the imputation process is performed in two stages. The first stage utilizes the predictive mean matching approach, where the actual values from the distribution are used as initial imputed values based on Euclidean distance. The second stage of the framework enforces the constraints within MICE and iterations enable us to get multiple optimal solutions for a given line item.

We tested our approach on a sample of stocks, where the missing values are scattered across several line items in the balance sheet, income statement and cash flow statement. The study reveals that the framework is able to capture the trend of specific variables, impute values for a few line items akin to their distribution, and yet solves the accounting constraints.

Application of multiple optimal solutions on the downstream model allows the practitioners to identify the parameters of interest in a scientific manner without eliminating the missing values. The approach will be useful for different groups such as academicians, analysts, investment managers for a variety of purposes such as studying default behavior, providing trading signals, investment opportunities, etc. Starting from pair trading based on pattern recognition to investment strategies of private equity, wealth funds and pension funds and asset allocation strategies of investment managers, fundamental data is the most important requirement to device any strategy. Banks can develop consumer analytics models for a portfolio that has limited data or missing data. Given the framework is developed more in a generalized manner, rigorous testing is recommended on downstream models, particularly out-of-time testing is performed in the research study.

The study intends to utilize the existing MICE framework and unravel the intricacies involved in implementing the imputation process on fundamental financial data. The major challenge while using the MICE framework on fundamental data is that the interdependencies among variables should be resolved simultaneously.

Despite the research framework being developed to handle missing values, the scope of the study can be extended further to include predictive and forecasting models. For example, if the top-most-level items are forecasted separately for a given company (such as Sales, and Assets) using this approach, other financial items can be easily derived by treating the other line items as a missing value problem. Secondly, the interrelationship between different items is not researched in detail in this approach. Using advanced methodologies such as random forests, the functional relationship for a given line item can be forecasted based on the predecessors of different companies with a similar asset size instead of using a single predecessor as a predictor variable. Thirdly, if the sales can be forecasted for benchmark companies (bell weather companies) using generic measures such as the Index of Industrial production and industry-wise predictors, the sales of mid-sized and small-sized companies can be forecasted using the tree approach developed in our framework. In this case, the sales of different mid-sized companies can be treated as a function of several benchmark companies. And the sales of small-sized companies can be derived using benchmark/mid-sized companies with their size as relative constraints. Fourth, the inter-relationship between different industries can also be easily explored. This can be achieved by combining industry-wide metrics with market-related information as predictors to forecast the sales of industry-wide companies. All the above problems can be solved by extending one of the three components that we have developed in this framework such as: i) solving the constraints; ii) improvising the functional form to predict a given line item; and iii) the top-down approach described as a tree.

7. References

[1] A. Pagano, D. Perrotta and S. Arsenis, "Imputation and outlier detection in banking datasets. In 46th scientific meeting of the italian statistical society.," 2012, May.

[2] D. B. Rubin and R. J. A. Little, "Multiple Imputaiton for Non-Response in Surveys," New York, 1987.

[3] P. Kofman and I. Sharpe, "Imputation Methods for Incomplete Dependent Variables in Finance," January 2000.

[4] Joos et al., "Credit classification: A comparison of logit models and decision trees. Proceedings Notes of the Workshop on Application of Machine Learning and Data Mining

in Finance," in *10th European Conference on Machine Learning*, Chemnitz (Germany), 1998b.

[5] D. J. Fogarty, "Multiple Imputation as a Missing Data Approach to Reject Inference on Consumer Credit Scoring," 2006.

[6] B. Galler and U. Kehral, "Missing Data Methods in Credit Risk," 2009.

[7] D. S. Bouhlila and F. Sellaouti, "Multiple imputation using Chained equations for missing data in TIMSS: a case study," 2013.

[8] S. Chen, "Jumping the hurdle of missing risk data," *Ambit Risk Institute,* 2013.

[9] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* pp. 1-38, 1978.

[10] D. Rubin, Multiple Imputation for Nonresponse in Surveys, 1987.

[11] D. Rubin, "Multiple imputation after 18+ years.," *Journal of the American statistical Association,* pp. 91(434), 473-489., 1996.

[12] I. R. White, P. Royston and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice.," *Statistics in medicine,* pp. 30(4), 377-399, 2011.

[13] H. Y. Zaslavsky, A. M. Landrum and ,. M. B. Harrington, "Multiple imputation in a large-scale complex survey: a practical guide.," *Statistical methods in medical research,* pp. 19(6), 653-670., 2009.

[14] E. A. Stuart, M. Azur, C. Frangakis and P. Leaf, "Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative," *American journal of epidemiology,* pp. 169(9), 1133-1139, 2009.

[15] R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 2002.

[16] T. De Waal, Handbook of statistical data editing and imputation, John Wiley & Sons, 2011.

[17] K. H. Li, X. L. Meng, T. E. Raghunathan and D. B. Rubin, "Significance levels from repeated p-values with multiply-imputed data.," *Statistica Sinica,* pp. 65-92, 1991.

[18] D. B. Rubin and N. Schenker, "Multiple imputation in health-are databases: An overview and some applications.," *Statistics in medicine,* pp. 10(4), 585-598., 1991.

[19] N. Schenker, D. J. Treiman and L. Weidman, "Analyses of public use decennial census data with multiply imputed industry and occupation codes.," *Applied Statistics,* pp. 545-556, 1993.

[20] S. V. Buuren, MICE: Multivariate Imputation by Chained Equations in R, 2007.

[21] T. E. Raghunathan, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey methodology ,* pp. 85-96, 2001.

[22] Buuren, et al., "Fully Conditional Specification in multivariate imputation," *Journal of Statistical Computation and Simulation,* pp. 1049-1064, December 2006.

[23] Kalb et al., "Mixtures of Tails in Clustered Automobile Claims," Monash University, Department of Econometrics and Business Statistics., 1995.

[24] N. Malhotra, "Analyzing marketing research data with incomplete information on the dependent variable," *Journal of Marketing Research,* pp. 74-84, 1987.

[25] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review.," *Journal of the Royal Statistical Society: Series A (Statistics in Society),* pp. 160(3), 523-541., 1997.

[26] M. Takahashi and T. Ito, "Multiple Imputaiton of Turnover in EDINET Data: Toward the improvement of imputaiton for the economic census" in *Work Session on Statistical Data Editing,* Oslo, Norway, September 2012.

[27] Tanner et al., "Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition," *Journal of the American Statistical Association,* pp. 953-960, 1996.

[28] King, Gary, et al, "List-wise deletion is evil: what to do about missing data in political science," *Annual Meeting of the American Political Science Association,* 1998.

[29] J. L. Schafer, Analysis of incomplete multivariate data, CRC press, 1997.

[30] J. L. Schafer, "Multiple imputation: a primer.," *Statistical methods in medical research,* pp. 8(1), 3-15., 1999.