World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:16, No:3, 2022

A Large Dataset Imputation Approach Applied to Country Conflict Prediction Data

Benjamin D. Leiby, Darryl K. Ahner

Abstract—This study demonstrates an alternative stochastic imputation approach for large datasets when preferred commercial packages struggle to iterate due to numerical problems. A large country conflict dataset motivates the search to impute missing values well over a common threshold of 20% missingness. The methodology capitalizes on correlation while using model residuals to provide the uncertainty in estimating unknown values. Examination of the methodology provides insight toward choosing linear or nonlinear modeling terms. Static tolerances common in most packages are replaced with tailorable tolerances that exploit residuals to fit each data element. The methodology evaluation includes observing computation time, model fit, and the comparison of known values to replaced values created through imputation. Overall, the country conflict dataset illustrates promise with modeling first-order interactions, while presenting a need for further refinement that mimics predictive mean matching.

Keywords—Correlation, country conflict, imputation, stochastic regression.

I. Introduction

IMPUTATION methods aim to estimate plausible values for gaps that may be found in datasets. Researchers have developed a large variety of methods to overcome missing values through imputation because imputation outperforms non-imputation methods and no single imputation method universally performs the best [1]. Rubin developed multiple imputation in the 1970s as a method for creating a value in a missing datum where uncertainty should be retained, and its remains the best general theory to deal with incomplete datasets [2]. The two main goals of multiple imputation are to estimate a value that is both unbiased and confidence valid [3]. However, some popular and preferred implementations of multiple imputation struggle to deal with datasets having a large number of data elements or datasets with high missingness. Van Buuren, a pioneer in multiple imputation by chained equations (MICE), lamented that large amounts of missing data or remotely connected data will influence the time required for convergence, where the key to convergence is to achieve independence in the imputations themselves [4]. Si agrees that multiple imputation faces operational challenges concerning their 409 variable large-scale dataset, explaining that MICE cannot directly handle skip patterns and requires additional efforts to account for logical or consistency bounds [5]. Others also contend that MICE is a superior approach in special cases, but faces problems with high-dimensional data [6], [7].

The motivating case study for this research uses data from the Internal Conflict Database, which is a repository of open-source data consolidated for the purposes of peace research. The open-source data comes from various data collectors such as the Center for Systemic Peace, the CIA World Factbook, Food and Agriculture Organization of the United Nations, Freedom House, World Bank, and a variety of other organizations. From the database, 932 continuous data proxies were selected representative of all aspects of society from political to economic to social themes in preparation of future region categorization research. The scope of the observations consists of the decade between 2006 to 2015, including the 173 United Nations (UN) member countries with over 250K total population as of 2016. Of the selected data elements, 74 capture complete data leaving the remaining vectors with an average missingness of 17.5%.

Prior country conflict research by Brantley [8] and Kane [9] demonstrated the superiority of MICE as the technique of choice for imputing missing data for country conflict data. Specifically, they both agreed that the multivariate method of predictive mean matching within MICE dominated other methods for most variables. Their assumptions rested on missing values being missing at random, which is made plausible by either limiting the country-year pair observations examined or limiting the scope of variables necessary for modeling. Brantley removed variables where entire country time-series periods were missing [8]. Kane chose only 32 significant variables from prior studies that predict country conflict, but only accounted for less than half of the percent missingness (6.79%) that is being researched in this study [9].

Attempting to apply their approach to a larger country conflict dataset resulted in algorithm computational failures. To illustrate, the R package MICE, used by both Brantley and Kane, failed to iterate one predictive mean matching pass of the 932 data elements within a 7-day computation period. Known barriers to algorithms like MICE include numerical problems from perfect prediction or collinearity, resulting in a failure to iterate [10]. A Python multiple imputation package, Iterative Imputer, also ran into computation issues, exceeding 64 GB of allocated memory after 15 iterations without converging.

In project management, it is often said that mangers must choose between only two of three constraints: time, cost, and quality. A similar sentiment may be said about analysis concerning time, computational power, and accuracy. With computational power being a fixed limiting constraint, a balancing act becomes necessary to implement an algorithm that maximizes accuracy within a reasonably

B. Leiby is with the Department of Operational Sciences, Air Force Institute of Technology, Wright-Patterson AFB, OH, 45433 USA (e-mail: benjamin.leiby@afit.edu).

D. Ahner is the Dean for Research, Air Force Institute of Technology.

defined time period. This paper presents an algorithm to impute very large datasets, outside the limits of existing packages, striking that balance between time and accuracy through a multiple imputation stepwise correlation multivariate regression approach.

The approach is similar to stochastic regression imputation where the point estimate from the regression equation is modified with a noise component to address upwards correlation bias and underestimated variability. Instead of relying on p-values to determine significant variables for the regression equation, a stepwise approach observing correlation values is presented to determine feasible independent variables where their significance is assessed through the increasing effect of the adjusted-R^2 statistic. This methodology development study, motivated by the country conflict dataset, imputes numerous variables without running into numerical problems.

II. Model Implementation

Rubin describes under a Bayesian approach that creating multiple sets of repeated plausible-imputed values reflects the uncertainty for the nonresponse when the procedure properly considers the complete-data estimates and the associated variance-covariance matrices [3]. That is, the estimates require an approach that considers errors on more correlated independent variables, rather than leaving some out, to overcome biased estimates and that combinations up to some level of interactions should possibly be considered [3]. The modeling approach used in this research takes advantage of a regression model with a noise component produced from the model residuals, also known as stochastic regression. Little views parametric models, such as regressions, as a strength in imputation as the assumptions are explicit [7]. Van Buuren demonstrated that the approach provides unbiased coefficients, although the coverage for confidence validity is not as good (0.908 vs 0.951/0.941) as more computationally intensive Bayesian and bootstrap approaches [2]. However, these computationally intensive methods like MICE become overly burdensome for imputing large datasets as discussed concerning numerical problems. With the regression approach, the benefit of using the residuals to incorporate the uncertainty in the imputation estimates rests on the assumption that the residuals are mean zero and normally distributed. The assumption was visually instantiated showing adequacy for both percent missingness and convergence rate as illustrated in Fig. 1.

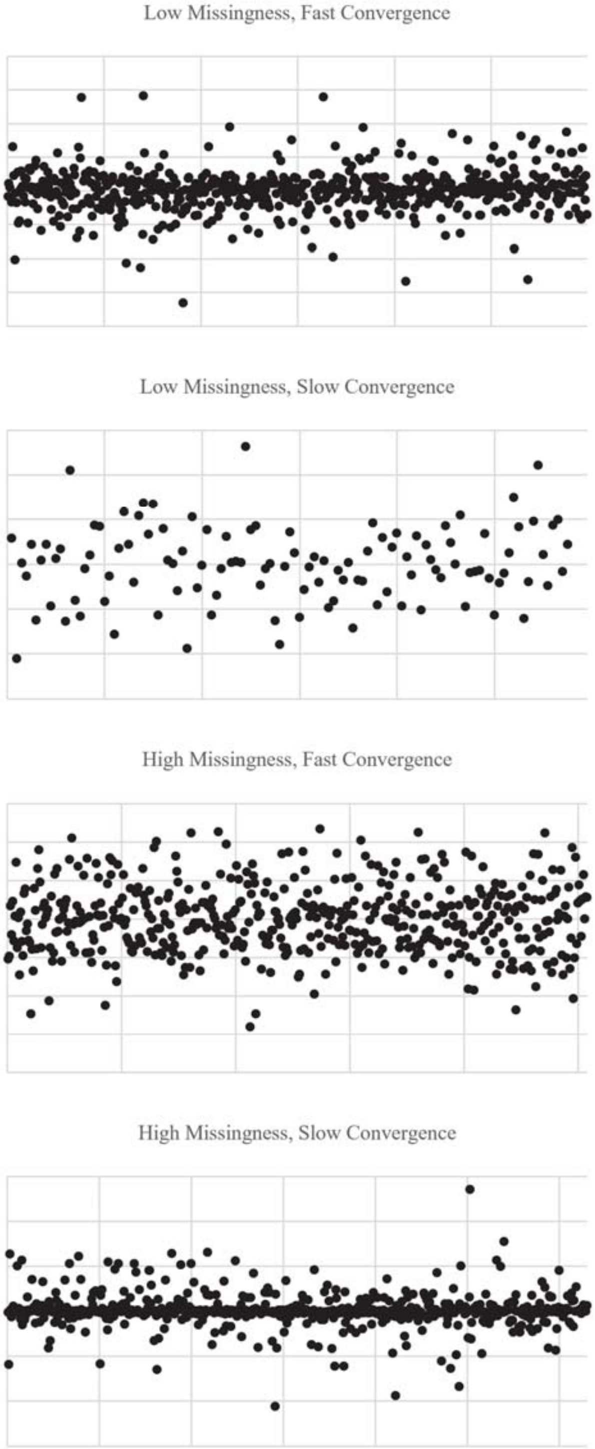


Fig. 1 Residuals for model regressions

The core component of the methodology resides in assuming correlated data elements should assist in providing accurate estimates for the missing values in the data. For example, height and weight are often seen as highly positive correlated variables, therefore if weight is missing in a few observations, it would be reasonable to use the height variable to impute the missing data points. Statistically, this concept is represented by the p-value, where the statistic is used to reject the null hypothesis that there is no relationship between the two variables. The benefit in starting with the analysis of correlation manifests in computation time. Whereas each variable would need p-value assessment in a stepwise regression for every iteration to determining significance, only one pairwise analysis of correlation coefficients is required to provide a static ordered list to assess significance for the entire dataset. The ordered list saves thousands of computations every iteration as the process is conducted once before model building rather than every time a model attempts to add a new variable.

Positive or negative correlation is inconsequential to the evaluation of the ordered list; the usefulness is that stronger relationships are considered first. The algorithm computes the absolute value of the Pearson correlation coefficients once, using only the known values in the dataset as seen in (1), where x_i and y_i are sample pairs in two different data elements with n non-missing value pairs. This matrix, **Q**, provides the foundation for discovering the strongest relationships that improve the model adjusted-R^2 within the least number of trials.

$$|r| = \left| \frac{n\sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{\sqrt{[n\sum_i^n x_i^2 - (\sum_i^n x_i)^2][n\sum_i^n y_i^2 - (\sum_i^n y_i)^2]}} \right| \quad (1)$$

Additionally, all data elements are rank ordered from the least proportion of missingness to the greatest proportion of missingness to identify the order in which the imputations will be processed. This ranking approach is similar to MICE where the least missingness is estimated first, in other words, optimizing the order of estimating the dependent variable within a regression model so subsequent imputations can benefit from observed and currently imputed values of all the other variables in the model [11]. The algorithm dynamically updates the dataset within each iteration to minimize estimation biases presented by missingness within the independent variables. That is, the approach assists in developing complete independent variables for subsequent imputation models, however, the model for the initial dependent variables may encounter missingness requiring preliminary simple imputation such as taking the mean. The biasing mean imputation on the independent missing variables is minimized by first imputing dependent variables with less missingness. As the dependent variable order processes the data elements with more missingness, the candidate independent variables become further complete with robust imputed values rather than weaker preliminary estimates to rectify their initial missingness. Furthermore, as the algorithm iterates, the bias decreases when the mean-estimated imputed independent variable becomes the dependent variable for imputation, garnering a better estimate from its own regression model. The rectification can be observed in the increased adjusted-R^2 for subsequently iterated models as seen in Fig. 2 and the quality of the normalized root mean square error discussed in the later sections.

Once these initial two processes of describing the **Q** matrix and dependent variable order are complete, the stepwise regression modeling commences. Using the data element missingness-related rank order, the data vector with the least missingness is set as the first dependent variable in need of imputation. Of the 932 data elements available, 74 already had complete data and did not require imputation, leaving 858 data vectors to impute. Using a stepwise approach, the algorithm adds independent variables to the model starting with the data element that has the strongest correlation according to matrix **Q** to the dependent variable.

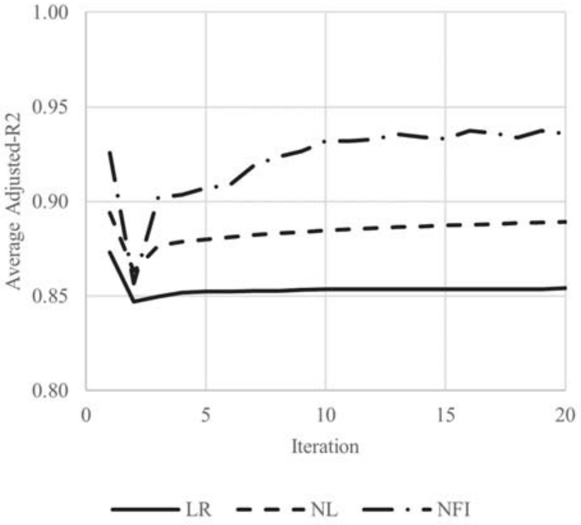


Fig. 2 Model average adjusted-R^2

While building the model, the algorithm sets aside a subset of complete data. For those instances when the dependent variable is missing, the associated independent variable data (observation) is removed from the subset. Furthermore, as additional independent variables are considered for inclusion into the model, initial cases arise where additional observations have missing or not yet imputed values in the set. These observations are also omitted from the subset. This mechanism of list deletion could potentially cause violations of the normality assumption of residuals if the degrees of freedom are too great with respect to the number of observations. Therefore, a threshold for an adequate number of observations was assessed before including the candidate independent variable.

There are a variety of recommendations to accommodate maintaining the normality assumption of residuals. For univariate regression, a general rule of thumb maintains at least 30 observations. For multivariant regression, 10-15 observations per independent variable has been demonstrated to be an optimal ratio [12]. A final strategy maintains to keep at least a quarter to half of the observations available for the most limited independent variable in the model. The most limited variable being the variable with the least number of known observations. Five thresholds were tested: 30 observations, 100 observations, and limiting variable observation ratios of a quarter, a third, and half.

The most conservative constraint (half of the most limiting variable) could reject the most plausible variable in the data set (highest correlation value) more often than desired inserting a less desirable variable concerning correlation value because it better suits maintaining the normality assumption of the residuals. Through testing, the most liberal constraint (at least 30 observations) was only enacted six times in the first iteration allowing the highest correlated variable to almost always enter the model, whereas the most conservative constraint forced an alternative 925 times.

No statistical difference at the 90% confidence level when observing the average, 25th percentile, or 10th percentile for the adjusted-R^2 of the model was identified, meaning model fit was not a factor. There was also no statistical difference when holding missingness as a factor. The most conservative constraint allowed some models to dip as low as 83 observations in the model dependent on the variables included, causing concerns about degrees of freedom and the normality assumption for a 10-variable model. Balancing maintaining a large number of observations while minimizing the number of alternative independent variables, the algorithm was set to a constraint of requiring 100 observations after listwise deleting missing values for model building. The selected cap of 10 independent variables corresponds to a minimum of 10 observations per variable, which is within the aforementioned optimal ratio. This constraint is only necessary for the first iteration as imputed values on subsequent iterations fill in any initial missing values in the data.

Next in the methodology, the candidate variable enters the model for adjusted-R^2 examination. The R^2 represents the explained variance by the independent variable toward the dependent variable. However, the R^2 continues to increase as more variables are introduced whereas adjusted-R^2 penalizes additional variables that fail to significantly affect the dependent variable. Three different models were examined: linear (LR), nonlinear (NL), and nonlinear with first-order interactions (NFI). The LR model, as illustrated in (2), provided the baseline case of providing parsimonious terms within the regression model, where y is the estimated dependent variable, x_n is the added known independent variable, β_0 is the model intercept, and β_{n1} is the corresponding linear coefficient. The assumption includes that any potential curvilinear relationships within the variables are insignificant. The NL model makes no such assumption and includes squared variable terms, in addition to the linear terms, as seen in (3), if those terms continue to increase the adjusted-R^2 of the model, where all coefficients from the linear model are present along with β_{n2} as the corresponding squared term coefficient. Additionally, the methodology observes the strong heredity assumption, that the geometric global extremum of all variables may not be the special case of zero [13]. The NFI model assesses both squared variables and first-order interactions, in additional to the linear terms, for inclusion as long as the adjusted-R^2 continues to increase for each candidate term as seen in (4) where all coefficients from the linear model are present along with β_{n3} as the corresponding interaction coefficient. Due to the assessment of each additional term, the computation times increases exponentially from LR to NL to NFI. Although the potential exists that additional variables may increase the adjusted-R^2 past 10 modeled variables, a cap of 10 variables was implemented. When considering country conflict datasets, Ray argues that country conflict data should adhere to Achen's "rule of three" when assessing independent variables for regression while Oneal demonstrates the rule to be too strict in examples of up to 8 variables [14]. Van Buuren notes that general regression, overcoming multicollinearity and degree of freedom problems, may be suitable upwards of 25 variables, however, explained variance after 15 variables is typically negligible at best [4]. The maximum 10 variables threshold facilitates a sweet spot to allow explained variance and manage the list deletion issue presented earlier.

$$y = \beta_0 + \beta_{11}x_1 + ... + \beta_{n1}x_n \tag{2}$$

$$y = (2) + \beta_{12}x_1^2 + ... + \beta_{n2}x_n^2 \tag{3}$$

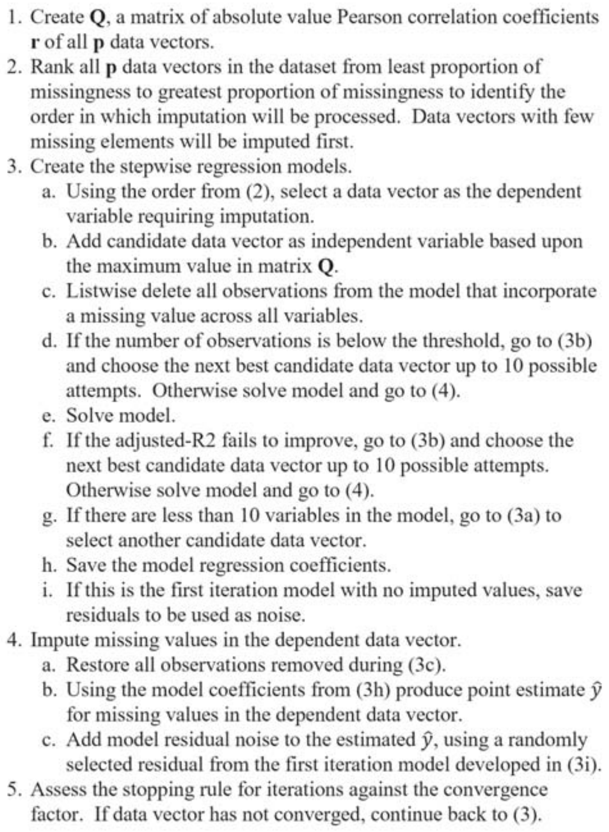
$$y = (2) + \beta_{n3}x_1x_1 + \beta_{n3}x_1x_2 ... + \beta_{n3}x_nx_n \tag{4}$$

Should the candidate variable fail to increase the adjusted-R^2, the next top 9 candidate variables are evaluated for inclusion. Observations concluded that on average, three initial candidates out of the 10 allowed variables in the model would fail to increase the adjusted-R^2, however an alternate variable was found to increase the adjusted-R^2 by the third best candidate, necessitating the need to look at subsequent independent variables past the initial failure to increase the adjusted-R^2.

Once the independent variables are identified for the model, the associated data produces the linear coefficients for the model that imputes the missing dependent values according to $\hat{y} = \boldsymbol{\beta} * \boldsymbol{X}$, where $\boldsymbol{\beta}$ are the model coefficient parameters and $\boldsymbol{X}$ are the data vector values for the associated missing dependent variable. This provides a point estimate from which to develop a stochastic regression result. For the first iteration, it is possible that some of the independent values may also be missing as discussed earlier, however, with trying to impute the dependent variable, list deletion is no longer an option. In these cases, an average of the non-missing data vectors estimates a feasible point estimate for the missing data. As previously mentioned, the bias inserted into the imputation diminishes with subsequent iterations as the dependent variable converges toward a more plausible value.

Noise added to the imputed point estimate provides the stochastic element desired in multiple imputation. Using a list of residuals captured from the first iteration, residuals produced only from the original known values, the imputed point estimate receives an adjustment from a randomly selected residual value to accommodate the uncertainty in the imputation. Seeing that the residuals are distribution normal, the uncertainty will have mean zero with standard deviation one.

Finally, the algorithm checks the stopping rule against the convergence factor to exit iterating each specific data element. The stopping rule compares each imputed before noise point estimate in the data vector from the before noise value of the previous iteration. Should all values within the data vector be less than the convergence factor, the algorithm considers the data element converged. For this study, each data element obtained a tailored convergence factor of three standard deviations of the data element's residuals to account for the different scale in values rather than rely on a static factor for the algorithm. The full pseudocode for the algorithm is provided in Fig. 3.



```
1. Create Q, a matrix of absolute value Pearson correlation coefficients
   r of all p data vectors.
2. Rank all p data vectors in the dataset from least proportion of
   missingness to greatest proportion of missingness to identify the
   order in which imputation will be processed. Data vectors with few
   missing elements will be imputed first.
3. Create the stepwise regression models.
   a. Using the order from (2), select a data vector as the dependent
      variable requiring imputation.
   b. Add candidate data vector as independent variable based upon
      the maximum value in matrix Q.
   c. Listwise delete all observations from the model that incorporate
      a missing value across all variables.
   d. If the number of observations is below the threshold, go to (3b)
      and choose the next best candidate data vector up to 10 possible
      attempts. Otherwise solve model and go to (4).
   e. Solve model.
   f. If the adjusted-R2 fails to improve, go to (3b) and choose the
      next best candidate data vector up to 10 possible attempts.
      Otherwise solve model and go to (4).
   g. If there are less than 10 variables in the model, go to (3a) to
      select another candidate data vector.
   h. Save the model regression coefficients.
   i. If this is the first iteration model with no imputed values, save
      residuals to be used as noise.
4. Impute missing values in the dependent data vector.
   a. Restore all observations removed during (3c).
   b. Using the model coefficients from (3h) produce point estimate $\hat{y}$
      for missing values in the dependent data vector.
   c. Add model residual noise to the estimated $\hat{y}$, using a randomly
      selected residual from the first iteration model developed in (3i).
5. Assess the stopping rule for iterations against the convergence
   factor. If data vector has not converged, continue back to (3).
```

Fig. 3 Methodology Pseudocode

III. Methodology Evaluation

The analyst trade-off of time, computational power and accuracy sparked the development of this methodology due to the "numerical problems" or "breakdowns" of the multiple imputation algorithm in alternative approaches. It is acknowledged that alternative approaches may foster improved plausible accuracy should the algorithms compile an iteration or process data in an acceptable period. This approach provides a choice to analysts with large datasets to balance acceptable time and accuracy. As General Patton suggested, "A good plan violently executed now is better than a perfect plan next week" [15]. In other words, this methodology allows analysts to have good imputations quickly instead of waiting for imputations from higher acclaimed algorithms that either may deliver too late or breakdown.

The time evaluation consists of observing the quantity of data elements converged after a certain number of iterations. Computationally, building the **X** matrix takes longer as the complexity of adding squares or interactions enter the model. Furthermore, looping back in the algorithm to find alternative independent variables increases iteration time as well. However, this time addition pales in comparison to the factor of how many data elements require imputation. Each data element takes 0.95 seconds to model under LR, 1.09 seconds under NL, and 1.68 seconds under NFI, with standard error in the milliseconds. The additional time for the more complex models is attributed to evaluating additional candidate terms, namely squared and first-order interaction terms. Recognizing that all models process data elements within a second of each other, the time component can be illustrated by how many data elements still require additional iterations to converge.

Preliminary model validation typically begins with assessing model fit by observing the dependent variable variability as a function of the independent variable variability known as the R^2 statistic. Good regression models desire independent variables that explain the variation in the dependent variable. The statistic is only useful if the residuals maintain the normal distribution assumption. Furthermore, the statistic always increases as additional independent variables are added to the model, therefore it has no stepwise assessment usefulness. The adjusted-R^2 penalizes additional variables allowing stepwise assessment. Observation of the adjusted-R^2 is twofold. First, a high value signifies that the imputations through the correlation approach may provide plausible values. Second, the initial observation of adjusted-R^2 contains only the known values in the original dataset. By the second iteration, bias was inserted into the dataset through estimating unknown values in the independent variables. Observing the adjusted-R^2 through subsequent iterations alleviates bias concerns as the value reapproaches the initial observation.

Finally, the normalized root means square error (NRMSE) functionally evaluates the goodness of the imputations to recreate known values. The NRMSE value is obtained by dividing the root mean square error by the range of the original data vector as illustrated in (5), where x_{1ip} are the known values in the test set, $\hat{x}_{1ip}$ are the imputed values corresponding to x_{1ip} with N_{1p} test set observations, x_{2p} are the known values in the original set, for the p^{th} data element of P total elements. Normalizing assists in adjusting the value to account for any scaling bias in the statistic with the common choice being range normalization [16]. A test set was created by randomly selecting 8% of the known data for imputation. Van Buuren stresses that imputation is a challenge "to obtain statistically valid inferences from incomplete data" rather than an exercise in accurately determining the unknown true value, especially when using multiple imputation techniques [2]. Despite his angst for root mean square error, he concedes that it is a good metric to evaluate the compromise between bias and variance if the desire is to assess accuracy and precision [2].

$$NRMSE = \sum_{p=1}^{P} \frac{\sqrt{\sum_{i=1}^{N_{1p}} (\hat{x}_{1ip} - x_{1ip})^2 / N_{1p}}}{\max(x_{2p}) - \min(x_{2p})} \tag{5}$$

IV. Model Results

The majority of data elements converged after only two iterations for the LR model and four iterations for NL. In other words, the difference between the regression point estimates in most vectors were less than three standard deviations of the first iteration residuals. The LR model converged more vectors faster than the other two as seen in Fig. 4; and with the fastest time to compute a data element, remained the fastest model type to reach the stopping condition.

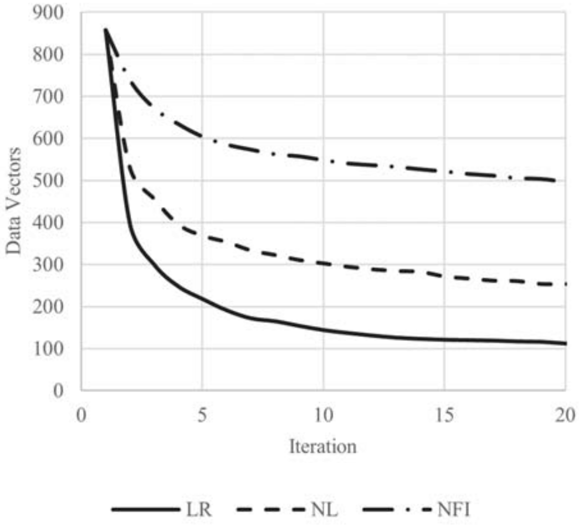


Fig. 4 Model convergence rate of data vectors

TABLE I
MODEL AVERAGE ADJUSTED-R^2, N=10

	Iteration 1		Iteration 20	
Model	Avg	Std Dev	Avg	Std Dev
LR	0.8732	0.0001	0.8541	0.0012
NL	0.8939	0.0000	0.8892	0.0037
NFI	0.9257	0.0003	0.9357	0.0035

The convergence rate appears counterintuitive when considering the average adjusted-R^2 of the models seen in Table I. It was hypothesized that better model fit would increase convergence, however, it was observed that the correlation between the convergence iteration and the data vector adjusted-R^2 was weak (<0.3). Despite this finding, all models produced a high average adjusted-R^2. With NL models producing a higher adjusted-R^2 than LR, the assumption remains plausible that many of the data elements should be characterized in curvilinear form. And supporting Rubin's claim, imputation models benefit further in adjusted-R^2 when modeling independent variables up to at least first-order interactions.

As far as the accuracy of the models, the median NRMSE for the data elements demonstrated low values after 20 iterations with 0.019 (LR), 0.0216 (NL), and 0.645 (NFI). However, the sum NRMSE was less optimistic with 1,903 (NL) and magnitude higher for NL and NFI. For the LR model, 4 of the 858 vectors had extremely high NRMSE values ranging from 11 to 1,154 inflating the overall NRMSE. All 4 vectors had very high adjusted-R^2 and no connection to percent missingness could be established. It was observed that some data vectors may have imputed values outside the plausible distribution. For example, known values in positive-only vectors had imputed observations with negative values. This remains an obstacle for regression methodologies that do not add limiting bounds like predictive mean matching. The "out-of-bounds" imputations exacerbate the issue for squared terms in the NL and NFI models when selected as independent variables, which lead to a larger number of outliers concerning vector NRMSE.

V. CONCLUSION

This paper presents a methodology to impute large datasets based on convergence of iterations within confidence bounds set by initial regression model residuals and using the information contained within the data correlation matrix. Large datasets increase the presence of numerical issues causing other imputation methods to fail. The regression methodology presented, demonstrated through the country conflict dataset, appears to overcome numerical issues without failed or stalled iterations. The methodology processes data elements quickly and generates high adjusted-R^2 models. Through developing the methodology, a stopping criterion to dynamically define convergence was presented offering a more tailorable condition for when data elements are of different scales. The exploitation of the initial regression model residuals overcomes any guesswork that may be present when submitting a static stopping tolerance offered in other imputation packages. The algorithm balances computation time, computational power, and accuracy to achieve a traceable, defensible approach to imputing large data sets where many preferred commercial packages fail. Despite the mentioned advantages, the methodology could benefit from further refinement. Although the methodology produces useable and defendable results, further work is needed to assure the user of plausible values. Notably, the issue of "out-of-bounds" imputations should be addressed to take further advantage of the improvements from NL and NFI type modeling. Other aspects of research could include investigating multicollinearities within the independent variables, while the dependent variable capitalizes on high correlation selection.

REFERENCES

[1] J. Luengo, S. García, and F. Herrera, "On the Choice of the Best Imputation Methods for Missing Values Considering Three Groups of Classification Methods," *Knowl. Inf. Syst.*, vol. 32, no. 1. 2012.

[2] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. CRC Press, 2018.

[3] D. B. Rubin, "Multiple Imputation after 18+ Years," *J. Am. Stat. Assoc.*, vol. 91, no. 434, pp. 473–489, Jun. 1996.

[4] S. van Buuren and K. Groothuis-Oudshoorn, "Multivariate Imputation by Chained Equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, Dec. 2011.

[5] Y. Si et al., "Multiple Imputation with Massive Data: An Application to the Panel Study of Income Dynamics," *arXiv Prepr. arXiv2007.03016*, Jul. 2020.

[6] Y. Deng, C. Chang, M. S. Ido, and Q. Long, "Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, Feb. 2016.

[7] R. J. Little, "On Algorithmic And Modeling Approaches To Imputation In Large Data Sets," *Stat. Sin.*, vol. 30, no. 4, pp. 1685–1696, Jan. 2020.

[8] D. Ahner and L. Brantley, "Finding the Fuel of the Arab Spring Fire: a Historical Data Analysis," *J. Def. Anal. Logist.*, vol. 2, no. 2, pp. 58–68, Jan. 2018.

[9] Z. J. Kane, "An Imputation Approach to Developing Alternative Futures of Country Conflict," Air Force Institute of Technology, 2019.

[10] C. D. Nguyen, J. B. Carlin, and K. J. Lee, "Practical Strategies for Handling Breakdown of Multiple Imputation Procedures," *Emerg. Themes Epidemiol.*, vol. 18, no. 1, pp. 1–8, Dec. 2021.

[11] C. O. Plumpton, T. Morris, D. A. Hughes, and I. R. White, "Multiple Imputation Of Multiple Multi-Item Scales When A Full Imputation Model Is Infeasible," *BMC Res. Notes*, vol. 9, no. 1, pp. 1–16, Dec. 2016.

[12] E. Núñez, E. W. Steyerberg, and J. Núñez, ”Regression Modeling Strategies”, *Rev. Española Cardiol*. (English Ed.), vol. 64, no. 6, pp. 501–507, Jun. 2011.

[13] J. A. Nelder, ”The Selection of Terms in Response-Surface Models—How Strong is the Weak-Heredity Principle?,” *Am. Stat.*, vol. 52, no. 4, pp. 315–318, May 1998.

[14] J. R. Oneal and B. Russett, ”Rule Of Three, Let It Be? When More Really Is Better,” *Confl. Manag. Peace Sci.*, vol. 22, no. 4, pp. 293–310, Sep. 2005.

[15] G. S. Patton and P. D. Harkins, *War As I Knew It*, Houghton Mifflin Company, 1995.

[16] Y. Luo, ”Evaluating The State Of The Art In Missing Data Imputation For Clinical Data,” *Brief. Bioinform.*, vol. 23, no. 1, Jan. 2022.