2-15-2023

Multicollinearity Applied Stepwise Stochastic Imputation: A Large Dataset Imputation through Correlation-based Regression

Benjamin D. Leiby

Darryl K. Ahner
Air Force Institute of Technology

Recommended Citation

Leiby, B.D., Ahner, D.K. Multicollinearity applied stepwise stochastic imputation: a large dataset imputation through correlation-based regression. J Big Data 10, 23 (2023). https://doi.org/10.1186/s40537-023-00698-4

Journal of Big Data

METHODOLOGY

Open Access

Multicollinearity applied stepwise stochastic imputation: a large dataset imputation through correlation-based regression

Benjamin D. Leiby* and Darryl K. Ahner

*Correspondence:
benjamin.leiby@afit.edu

Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433-7765, USA

Abstract

This paper presents a stochastic imputation approach for large datasets using a correlation selection methodology when preferred commercial packages struggle to iterate due to numerical problems. A variable range-based guard rail modification is proposed that benefits the convergence rate of data elements while simultaneously providing increased confidence in the plausibility of the imputations. A large country conflict dataset motivates the search to impute missing values well over a common threshold of 20% missingness. The Multicollinearity Applied Stepwise Stochastic imputation methodology (MASS-impute) capitalizes on correlation between variables within the dataset and uses model residuals to estimate unknown values. Examination of the methodology provides insight toward choosing linear or nonlinear modeling terms. Tailorable tolerances exploit residual information 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 methodology provides useable and defendable results in imputing missing elements of a country conflict dataset.

Keywords: Correlation, Country conflict, Imputation, Stochastic regression

Introduction

Many popular multiple imputation methods rely on a regression framework to develop plausible missing values [1]. Although no single imputation method succeeds at being the best in all imputation applications [2], some studies demonstrate k-nearest neighbors as the best single imputation method and predictive mean matching (pmm) as the best multiple imputation method for the datasets considered [3]. Prior country conflict dataset imputations by Ahner and Brantley [4] and Kane [5] also contend that pmm, a regression approach multiple imputation bounded to only known values for estimates, exhibited superior performance toward their country conflict datasets compared to other tested approaches. However, these prior studies were limited to small datasets of 32 variables. When expanding the 32-variable country conflict dataset into a very large dataset, the preferred pmm approach broke down due to numerical problems [6]. In

Leiby and Ahner [6], a new regression approach investigated capitalizing on dependent variable correlation in a stochastic regression framework to overcome numerical problems. Although the approach provided promise with favorable results, the algorithm also suffered from some "out-of-bounds" imputations and concerns over multicollinearities within the independent variables [6]. This research extends the Large Dataset Imputation through Correlation-based Regression approach found in [6] to develop robust imputations by including variable range-based guard rails and exploring correlation selection discounts.

The large dataset considered consists of 932 continuous data proxies or data elements from the Internal Conflict Database [6] allowing direct comparisons between the initial method's results and the extension presented in this paper. The scope of observations involves annual data over 10 years from 173 United Nations (UN) member countries that possess a total population of over 250K. The observations are recorded as country-year pairs for a total of 1730 observations. This dataset supplies a diverse selection of multiple data elements spanning all three country conflict aspects of political, economic, and social influences. Completing the dataset with plausible imputations assists peace researchers in developing solutions through increasing sample size power, especially when employing analytical modeling.

Within the dataset, 74 of the 932 data elements were complete cases with all country-year observations. Considering all three patterns of missingness, the missingness of an observation in a data element averaged 17.5% while the missingness of a data element for a country-year pair observation averaged 14.0%. The diversity of missingness in this large dataset presents a good opportunity for multiple imputation which has demonstrated to be robust even when datasets depart from the normality assumption or when the proportion of missingness may be high [7].

The overwhelming majority of data analysis techniques require complete data as mathematical operations cannot be applied to non-values. An easy solution to overcoming this problem is using listwise deletion on the observations with missing values, however, such methods increase biasness, underpower sample sizes, or insert unreliable estimates [8]. For example, when considering this dataset, listwise deletion would reduce the desired 1730 observations down to an unacceptable 3. Imputation, a mathematical process of inferring a value to an undocumented attribute of an observation, is then necessary to create a more useable dataset for analysis.

With Rubin's proposal of multiple imputation, the identification of three distinct patterns of missingness became standard practice, which include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [9]. Within the large dataset, all three categorizations can be observed, which accounts for some of the numerical problems encountered when applying pmm. Country conflict data often carries the complexity of missing data through multiple lenses: seeing missingness through unique country-year pairs as observations, missingness through unique countries over a time-series of years, and missingness as individual occurrences across multiple variables. Some of the missingness could be identified as missing at random, while some are obviously worse case as missing not at random.

MCAR is data missing as a random effect in the sample, or in more colloquial terms, due to just bad luck. The missingness is not correlated or dependent to any observed

or unobserved independent or dependent measurement. MCAR data is rarely found in practice, however, it can be perceived that very low missingness in a dataset could be identified as such. Each complete data vector would consist of 1730 observations, where each represented country may have 10 time-series data points. A worst-case scenario would imply that all missingness came from one country. Therefore, keeping missingness less than half the country observations and claiming MCAR would place a data element with at most 4 missing observations and still be considered MCAR. This happens in 73 of the 932 data elements.

More commonly, MAR ties the missingness to an observed measurement, however, the missing data does not depend on the value of the missing data. In the dataset, such missingness may manifest in areas such as the *Corruption Perception Index* score not being recorded for a country that is in conflict or data not being measured due to a country having an autocratic government and controlling what information is available to the public. However, the missing values may be validly imputed by considering other observed variables in a model. This assumption would fit the majority of missingness in the dataset.

MNAR ties the value of the missingness to the missing value itself or when the missingness may not be understood by any other observed value. This could manifest at the intersection where both high missingness rates are seen across the time-series and within a variable column. Such examples include a country having no time-series data for a variable and the variable across countries also having high missingness; for instance, observational data for the Democratic Republic of Korea having no time-series data for *Battle-Related Deaths* along with the data element also having a cumulative 84% missingness. This applies to at least three data elements which are observed with scrutiny.

Two main issues surfaced in [6] while developing the concept for the Large Dataset Imputation through Correlation-based Regression approach. First, there are concerns about multicollinearity of independent variables effecting the stability of regression coefficients. Second, regression results may produce imputation estimates that are outliers to the distribution of known values undermining the confidence in the plausibility of the imputed values. The combination of these two issues are assumed responsible for the imputed data vectors that experienced extremely high root mean square error values [6]. These issues are addressed in this new imputation process Multicollinearity Applied Stepwise Stochastic Imputation (MASS-impute).

Model implementation

Reviewing the original proposed algorithm, the steps can be categorized into three main segments: pre-processing, regression modeling, and imputation development. Pre-processing consists of two parts: developing a correlation matrix used for nominating variables and a ranking of variables by missingness. The correlation matrix consists of the absolute value Pearson correlation coefficients, $\boldsymbol{r}$, used for variable selection, and designated as matrix $\mathbf{Q}$. The rank ordering of data elements establishes that order for imputation with the least missing elements undergoing the imputation process before data vectors with more missingness. This is consistent with other imputation methods using

multiple imputation by chain equations (MICE) [10]. No changes to the pre-processing segment were made from the original method in [6].

The modeling segment selects up to 10 variables for inclusion into a regression model to estimate the missing values of a single data element, of which 96% of data elements typically select the maximum, varying slightly from iteration to iteration. Producing candidate regression coefficients uses a stepwise process of evaluating candidate variables with the goal of increasing the adjusted-R^2 statistic. The original method selected candidate independent variables that had high r linear correlation scores with the dependent variable. The method structures itself by leveraging variables that provide as much useful information as possible to estimate missing data points. Theoretically, if one independent variable were perfectly correlated with the dependent data element having missing values, then it would be expected that perfect prediction could be obtained. Therefore, when selecting individual candidate variables for inclusion into the model, the main criteria focuses on increasing the adjusted-R^2 without regard to multicollinearity with other independent variables. Often, analysts highlight multicollinearity as a concern when building models; modeling with highly correlated independent variables produce unstable estimates, inflated variances, and confounding effects, although coefficient instability may be a consequence of multicollinearity rather than a product of it [11, 12]. To clarify, the perceived multicollinearity problem consists between only the univariate independent data elements themselves but not with modeling constructs such as interaction product terms. Modeling square terms or product terms often highly correlate with the individual independent variables, yet do not create multicollinearity problems as "multicollinearity neither affects the value of the coefficient of the product term nor inflates its standard error" [13]. The multicollinearity problem typically concerns model analysis rather than modeling for imputation purposes. Still, imputation practitioners pause for concern when reading van Buuren's statement that using several hundred variables in multiple imputation cannot be feasible due to multicollinearity and computational problems [14]. Solutions to the multicollinearity problem often include removing variables to increase parsimony. Some suggest removing variables in imputation models should they have large amounts of missing data due to incomplete cases, failure to have adequate association with the dependent variable (absolute correlation value greater than 0.5), or high correlation with other independent variables resulting in not adding additional value to the model [12]. Yet excluding variables with high partial correlation simultaneously increases the risk of omitted variable bias [11]. Despite the hazards of multicollinearity, the implications may be better described as a problem of degree rather than kind [15], therefore this research presents variable selection conditioned on degrees through correlation discounting. In other words, multicollinearity of independent variables may be addressed through variable selection with a discount.

Before applying a discount to the variable selection criteria, it is necessary to establish when to apply a discount. If correlation is too high, multicollinearity concerns exist and discounting is deemed necessary. If discounting is applied too heavily, the algorithm may omit valuable variables resulting in a less than optimal imputation. To aid in proper variable selection, five categories of correlation are defined: very high

Table 1 Correlation categories with no discounting

Correlation category	Number of data elements	Percent of elements including category (%)
Very high (1.0–0.9)	597	64
High (0.9–0.7)	709	76
Moderate (0.7–0.5)	829	89
Low (0.5–0.3)	919	99
Negligible (0.3–0.0)	932	100

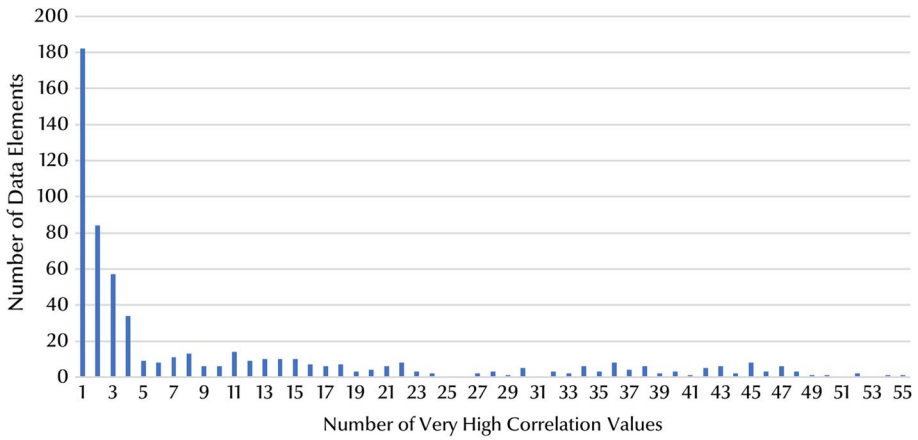

Fig. 1 "Very high" correlated values in data elements

(1.0–0.9), high (0.9–0.7), moderate (0.7–0.5), low (0.5–0.3), and negligible (0.3–0.0) as illustrated in Table 1. Some data elements have correlation values in each correlation category while others data elements may only be represented in a few categories. It was noted that 5 of the 932 data elements consisted of all correlation values below 0.5, suggesting they would not be strong candidates for inclusion in the model.

The method uses a forward stepwise linear regression approach, in the form of $\hat{y} = \boldsymbol{\beta} * \boldsymbol{X}$ where $\hat{y}$ are imputed results from $\mathbf{X}$ data elements with associated $\boldsymbol{\beta}$ coefficients, which economizes on computational effort. Through this method of stepwise addition by correlation value, the method nominates variables with the highest r absolute value. Limiting the multivariate regression equation to only 10 variables, it is highly unlikely that correlation values below 0.5 are included in any models. However, addressing multicollinearity, the high correlation between independent variables adding little value is addressed through exploring discounting of a variable's r value based on its correlation with variables already in the model. This provides the first deviation from [6] as variables are now nominated through a discount matrix rather than matrix $\mathbf{Q}$ in order to mitigate multicollinearity between independent variables. Data elements with at least one very high correlated variable account for 64% of the dataset, having a median number of just one variable, as illustrated in Fig. 1. If no discounting is present when selecting variables, 415 data elements have at least the first two candidate variables with an absolute value collinearity above 0.9 and 139 data elements potentially consisting of all 10 variables within that very high category. The discount process alleviates this situation.

Four discount strategies were examined on a degree scale and are described as follows: None, Cube, Square, and Max. The None discount is the base case where no discount is applied. The algorithm chooses the next best variable based on correlation with the dependent variable. This baseline case illustrates the effects of multicollinearity among independent variables in the imputation model and whether multicollinearity should be a concern. Although almost-linear related predictors are frequently a source of problem for imputation [8], this baseline assists in quantifying how much a problem may be present within the large country conflict dataset considered [14]. At the other extreme is the Max discount. The Max discount chooses the next best variable based on adjusted correlation with the dependent variable by comparing each candidate variable's correlation with the dependent variable after subtracting the maximum correlation between the candidate variable and the variables already included in the model. The Max discount, along with the Square and Cube variant can be seen in Eqs. 1–3 (Max, Square, Cube respectively), where $\mathbf{A}_{i,j}$ is the discounted absolute value correlation score, $\mathbf{Q}_{i,0j}$ are the original absolute value Pearson correlation coefficients for dependent variable i and nominated independent variable j, and $\mathbf{Q}_{i,nj}$ are the original absolute correlation values of variables currently added to the model associated with the dependent variable. Matrix $\mathbf{A}$ then, in all cases, is the transformed correlation matrix after the appropriate discount from with to choose the next data element j with the maximum discount value.

$$Max : \mathbf{A}_{i,j} = \mathbf{Q}_{i,0j} - max(\mathbf{Q}_{i,1j}, \mathbf{Q}_{i,2j}, \ldots, \mathbf{Q}_{i,nj}) \tag{1}$$

$$Square : \mathbf{A}_{i,j} = \mathbf{Q}_{i,0j} - [max(\mathbf{Q}_{i,1j}, \mathbf{Q}_{i,2j}, \ldots, \mathbf{Q}_{i,nj})]^2 \tag{2}$$

$$Cube : \mathbf{A}_{i,j} = \mathbf{Q}_{i,0j} - [max(\mathbf{Q}_{i,1j}, \mathbf{Q}_{i,2j}, \ldots, \mathbf{Q}_{i,nj})]^3 \tag{3}$$

For example, using Max discount, consider the data *a–f* as shown in Table 2. Data element *a* is set as the dependent variable and data elements *b* and *c* as independent variables already in the model. For every unmodelled data element, elements *d–f*, subtract from the associated correlation value of *a*, the maximum value between the correlation values associated to the already modeled data elements *b* and *c*. Data element *e* would be selected for the next modeled independent variable having the highest adjusted correlation value after discount. The Square and Cube discounts choose the next best variable based on adjusted correlation with the dependent variable after subtracting the respective squared maximum or cubed maximum absolute value correlation value of each currently modeled term from the dependent variable's correlation value, thus reducing the effect of the discount. Additionally, the adjusted correlation value of the candidate data

Table 2 Adjusted correlation using max discount

Unmodeled variable	Correlation with a	Correlation with b	Correlation with c	Discount (max.)	Adjusted correlation
d	0.8	0.1	0.7	0.7	0.1
e	0.5	0.2	0.1	0.2	0.3
f	0.4	0.1	0.2	0.2	0.2

element must have a positive value, $\mathbf{A}_{i,j} \geq 0$, or the algorithm stops adding variables to the model. All values in $\mathbf{Q}$ are absolute values, so a negative discount value would be an imaginary number and infeasible for consideration.

Through the discounting, all data elements eliminate any second candidate variables having very high collinearity as seen in Table 3. The quantity of data elements potentially choosing a second candidate variable with high collinearity is noted under the column quantity of data elements. However, by the third selection of a candidate variable, all variables would be in the moderate category thus satiating any concerns about multicollinearity, but potentially increasing the risk of omitting key variables. Comparing the validation statistics between the degrees of discounting should identify where the balance may lie between too much collinearity and key variable omittance.

Now that the degree of multicollinearity is addressed, the implementation of preserving the stochastic element of imputation is modeled. Within the modeling segment, the stochastic noise values are saved. For the first iteration of the algorithm, only known values (non-imputed) are used to select variables and produce coefficients. The residuals from this first iteration are saved and set aside to be used for all subsequent iterations as stochastic variation as well as in determining convergence.

The final segment, imputation, takes the regression coefficients from modeling, applies them to the related independent missing data values, and produces a point estimate. Randomly choosing a value from the normal distribution of residuals saved on the first iteration provides the uncertainty added to the estimate. The second modification to the original algorithm concerns setting limiting bounds for estimating imputations. The original regression models were unbounded and therefore could produce unreasonable estimates unlike a contrasting methodology such as pmm. Taking the *Battle-Related Deaths* data element as an example, the model regression coefficients could estimate some imputed observations with negative numbers. A negative death has no clear or rational interpretation, which implies that the data element should not allow for such values. Additionally, there are no known negative values in the original data distribution, which would cause further plausibility concerns if left unchecked. Therefore, the imputed estimates are assessed with consideration toward the known values within the data vector. This assessment acts like guard rails. There are three types of variable range-based guard rails implemented in the algorithm. First, if the minimum and maximum of the known data points are 0 and 100, it is assumed the data vector is a percentage and therefore all imputations are bounded between 0 and 100. Second, if the known data points present no negative values, it is assumed the data must be positive only and bounded as such. Third, if

Table 3 Second variable correlation categories after discounts

Correlation category	Number of data elements			Percent including category		
	Cube	Square	Max.	Cube (%)	Square (%)	Max. (%)
High (0.9–0.7)	25	8	3	3	1	0
Moderate (0.7–0.6)	82	41	7	9	4	1
Moderate (0.6–0.5)	230	98	28	25	11	3
Low (0.5–0.3)	902	244	153	97	26	16

the known data vector contains both positive and negative values, then the bound set is 1.5x the maximum and minimum known values. The wider range accounts for potential unobserved nonresponses outside the observed values in the model without allowing extreme extrapolation. Any imputed point estimates that are outside the bounds are set to the bound and then applied with applicable noise to stay within the bounds. However, some point estimates that are already within the bounds still may produce imputations outside the bounds when the stochastic element of noise is applied, therefore, the data vector is assessed a second time after the noise application to ensure all imputed values remain inside the bounds.

The final step in the imputation segment considers the stopping condition. Due to the importance to the process, the idea of convergence is expounded. One of the largest issues plaguing multiple imputation techniques manifests in knowing when enough iterations are complete. Defining convergence becomes even more of a nebulous term because of the stochastic nature of the algorithm accounting for the uncertainty of the imputed value. Stochastic convergence has four main definitions: observing a convergence in distribution, a convergence in probability, a convergence almost surely, and convergence in r-mean. Van Buuren notes that there is no clear-cut method for determining convergence in multiple imputation, however, the MICE package in R defines convergence as "when the variance between the different sequences is no larger than the variance with each individual sequence" [14]. A Python implementation of MICE in Iterative Imputer notes that their experimental algorithm could warrant more investigation into their convergence criteria (#14338) where certain datasets fail to converge and debate continues on what criteria to use against the tolerance parameter [16]. The Autoimpute documentation does not expound upon stopping conditions and settles with simply stating that increasing the posterior sampling chains may improve the chance of convergence [17]. Nevertheless, an algorithm benefits from a stopping condition to assess the completion of the imputation outside a user defined value for iterations, which typically is convergence within a tolerance.

The noise aspect in stochastic regression adds uncertainty to the regression point estimate by exploiting the residuals in the known data points. Leveraging van Burren and the sentiments expressed in Iterative Imputer and Autoimpute, consecutive iterations of dependent variables within the distribution range of the residuals should satisfy a classification of convergence. The difference between the regression point estimate and its prior iteration estimate becomes the assessment for convergence. These estimates are prior to the addition of the stochastic noise. If every observation in the data element for the iteration has an absolute value difference less than the stopping criteria, then the data element is converged and no longer assessed for imputation. This leads toward the question of a good stopping criteria. In the three previously mentioned commercial programs, the stopping criteria is a user inputted tolerance. However, a user inputted tolerance does not account for the different scales that may be present in the large set of data elements. Capitalizing on the residuals used for the stochastic nature of the algorithm can assist in formulating tailored stopping tolerances for each data element. The return on a tailored tolerance manifests in observing the distribution of the first iteration residuals for each data vector. Observing the adjusted-R^2, experimenting with various

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 is processed. Data vectors with few missing elements are 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. Create matrix **A** using the discount strategy.
    c. Add candidate data vector as independent variable based upon the maximum value in matrix **A**.
    d. Listwise delete all observations from the model that incorporate a missing value across all variables.
    e. If the number of observations is below the threshold, go to (3c) and choose the next best candidate data vector up to 10 possible attempts. Otherwise solve model and go to (4).
    f. Solve model.
    g. If the adjusted-R^2 fails to improve, go to (3c) and choose the next best candidate data vector up to 10 possible attempts. Otherwise solve model and go to (4).
    h. If there are less than 10 variables in the model, go to (3b) to update matrix **A** and select another candidate data vector.
    i. Save the model regression coefficients.
    j. 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 (3d).
    b. Using the model coefficients from (3i) produce point estimate $\hat{y}$ for missing values in the dependent data vector.
    c. Using variable range-based guard rails, ensure all point estimates are plausible.
    d. Add model residual noise to the estimated $\hat{y}$, using a randomly selected residual from the first iteration model developed in (3j).
    e. Recheck guard rails to ensure all imputed estimates are plausible.
5. Assess the stopping rule for iterations against the convergence factor. If data vector has not converged, continue back to (3).

Fig. 2 Multicollinearity applied stepwise stochastic imputation (MASS-impute)

standard deviation tolerances of the residuals, little improvement manifests in selecting a tighter than three standard deviation parameter for the stopping condition tolerance.

Acknowledging the initial presentation of the algorithm presented in [6], the modifications to the pseudocode are presented in Fig. 2. The inclusion of the discount strategy is seen in step 3b with additional effects in 3c and 3h. Instead of eliciting candidate variables from the **Q** matrix, candidate variables are selected from the **A** matrix, which is updated with every variable selection. The variable range-based guard rails are introduced in step 4c with a second variable range-based guard rail check in 4e.

Methodology evaluation

Validation of the methodology continues with the same three metrics conducted in [6]: time evaluation illustrated by number of data element convergences, model fit calculated by adjusted-R^2, and prediction accuracy through the proxy of recreating known values through imputations and assessed under a normalized root mean square error (NRMSE). Due to different scales between the data elements, normalization of the error is necessary to make comparisons between data elements. The normalization used in this study leverages the original range of the data vector

as illustrated in Eq. 4, 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, and x_{2p} are the known values in the original set, all for the pth data element of P total elements. The test set randomly selected 8% of known observations to be recreated through imputation. Finally, since an instance of an imputed value is not unique, 30 imputed complete datasets are used for analysis.

$$NRMSE = \sum_{p=1}^{P} \frac{\sqrt{\sum_{i=1}^{N_{1p}} (\hat{x}_{1ip} - x_{1ip})^2 / N_{1p}}}{\max(x_{2p}) - \min(x_{2p})} \tag{4}$$

Randomly removing known values and checking the accuracy of the imputations against the known values provides an evaluation akin to MCAR. When data is MAR, other validation measures may be more appropriate. Van Buuren observed distributions and scatterplot values to observe if the estimates overlayed with known results appeared as if nothing had ever been missing when checking the plausibility of multiple imputation results [18]. For a set of imputation results, visual inspection via a scatterplot should present further evidence about the plausibility of the imputation, both in distribution and in position.

There are also a few statistical tests to evaluate the plausibility of results. Should the known values follow a normal distribution, or the quantity of imputed values be sufficiently small, a parametric two-sample t-test would highlight inconsistencies in the means. The null hypothesis being that the known data and the imputed data are drawn from populations that share the same mean. If the p-value of the test is greater than some confidence level, then the difference in means appears insignificant and the perception is that the sample means are the same and assumed to come from similar distributions. However, if the distribution is unknown, the non-parametric Wilcoxon–Mann–Whitney (WMW) test also highlights inconsistencies between two independent groups, but with relation to medians. WMW test asserts that if the data values of two quantities x_n and y_m are ordered, the arrangement when counting how many times y precedes x, designated as U, is significant if $P(U \leq \overline{U})$ is under some confidence interval [19]. The null hypothesis states that the known data and the imputed data are drawn from populations that share the same median. These two inferential tests examine descriptive metrics of the imputations; therefore, a goodness-of-fit test is also examined. Two well-known goodness-of-fit tests are the Kolmogorov–Smirnov and the Anderson–Darling. A simple understanding of the two tests see the one-sample Kolmogorov–Smirnov test as a supremum proximity analysis of the empirical distribution function, and the one-sample Anderson–Darling test as an evaluation of how close the points are to a straight line estimated in a probability graphic [20]. The two-sample Anderson–Darling (AD) test is similar to the Kolmogorov–Smirnov in that it is a goodness-of-fit test, but is said to dominate Kolmogorov–Smirnov in observing smaller moments in the distribution [4] due to its sensitivities in the extreme ends of distributions [21]. For this research, the AD is used, consistent with other country conflict imputation research [4, 5]. The null hypothesis proposes that the known and imputed values are drawn from the same population without having to specify the distribution function of that population.

Python SciPy packages [22] provided ease of use implementation to generate p-values. The ttest_ind package set the assumed variance between the vectors as not equal. The mannwhitneyu and the anderson_ksamp packages used default values. This study used a significance level of 95%. Each imputed dataset was assessed against the known values in the data element vector to quantify how many imputation sets satisfied the test.

Model results

The model results highlight the benefits of the methodology in three aspects: micro, macro, and comparative. The micro aspect looks at the application of variable range-based guard rails, a change in controlling the aperture of the results, with a focus on improving the methodology to the previous evolution. The macro aspect evaluates the application of discounting, a change in nominating variables for inclusion, with a focus on identifying the degree of multicollinearity hindrances and objectively selecting the optimal discount. The comparative aspect dives into the imputations themselves when the method is optimally configured to defend the plausibility of the method's results.

When researching categories of correlation, Nguyen highlighted that independent variables with inadequate association toward the dependent variable should be removed from the model [12]. Five data elements had a maximum correlation value below 0.5, which, using Nguyen's advice, would recommend no modeling variables for imputation. Of the five data elements containing only correlation values below 0.5, their percent missingness were 0.3%, 2.1%, 5.2%, 7.9% and 36.2%. Despite their correlation limitation, the three lowest missingness met the convergence criteria in all model-runs by at least iteration 8 and therefore should not be a cause for concern for instability. The two with higher missingness would often converge by iteration 5, although 3 of 10 exploratory model-runs saw non-convergence when allowed to run out to 100 iterations. Still, the data elements below the minimum threshold by Nguyen do not appear to unduly suffer regarding the validation metrics within this methodology and therefore it is likely that Nguyen's bottom threshold of 0.5 may be set too high.

Micro aspect

The application of variable range-based guard rails provided many benefits to the models. As in [6], three regression model constructs were considered, linear (LR), nonlinear (NL), and nonlinear with first-order interactions (NFI). The LR model retained the fastest convergence rates compared against the NL and NFI models, and the 'with variable ranged-based guard rails (WGR) continued to improve all models compared to the original (Orig) models from [6] as illustrated in Fig. 3. The overall time comparisons between the Orig models compared to the WGR models is less pronounced, although the NL-WGR model converged faster than the LR-Orig model. In practice, the LR-Orig model completed 20 iterations after 52.3 ± 0.4 minutes using an Intel i7-9700K with 64GB of RAM in Python 3.8.8, however, even with the inclusion of the variable range-based guard rails increasing the checks within the algorithm, increasing the complexity of the models with squared terms still enjoys similar completion times due to converging on earlier iterations. This is significant where each additional unconverged data element adds compounding time to the completion of an iteration where the NFI-Orig model finished after 3 h 50 min for an average of

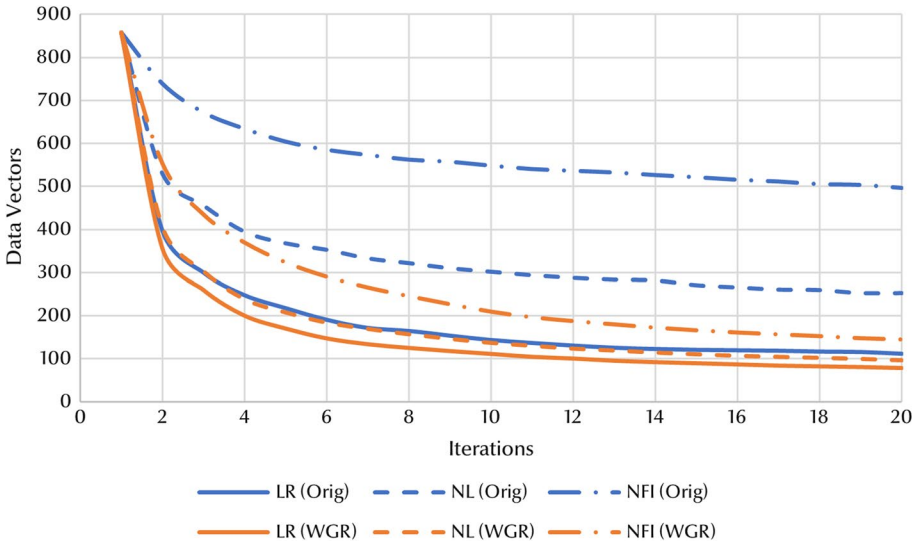

Fig. 3 Model convergence rate of data vectors, N = 10

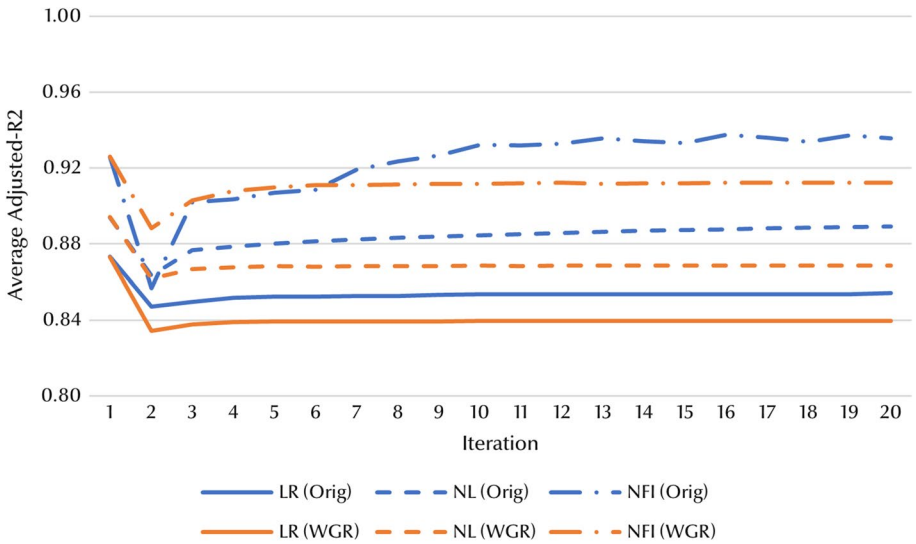

Fig. 4 Model average adjusted-R^2, N = 10

461 unconverged data elements. The algorithm with variable range-based guard rails included more statistical outputs, so a direct comparison is not practical, but even with the additional workload, the NFI-WGR finished significantly faster with an average of 2 h 49 min. In fact, the NFI-WGR averaged only 33 unconverged data elements more than the LR-Orig model, converging sooner, and therefore theoretically outputting faster than NL-Orig or NFI-Orig models.

The model fit continued to retain similar features with or without variable range-based guard rails. As seen in Fig. 4, the second iteration saw a decrease in adjusted-R^2, which is an artifact of both a preliminary mean imputation for missing values in the independent variables for only iteration 1 as well as the iteration 2 models being

constructed with more observations from the first round of imputations. As with the Orig models, this one-time mean imputation bias decreases as each round of imputations develops more plausible results and converges on a value within the range of noise. Although the WGR models do not rebound to the level of the Orig models, the measurement retains average values above 80% and demonstrate more stable results, especially when implementing the NFI model.

The main benefit of the variable range-based guard rails surfaces when cross-checking known values against imputed values. Although the median NRMSE values of the Orig models demonstrated low values, the LR-Orig model sum value was quite high due to four outliers and the NL-Orig and NFI-Orig sums were excessive due to compounding artifacts of outliers. When variable range-based guard rails are implemented, these outliers are severely reduced. The maximum data element NRMSE were 0.471 (LR-WGR), 0.477 (NL-WGR), and 0.417 (NFI-WGR) with variable range-based guard rails as opposed to values from the Orig models without variable range-based guard rails that ranged into the thousands. As with results from the Orig models, the distributions are still not normal as shown by averages of 0.054 (LR-WGR), 0.049 (NL-WGR) and 0.042 (NFI-WGR), and median values lower at 0.022, 0.019 and 0.013 respectively. This brings the NRMSE sums into reporting range: 50.120 ± 0.059 (LR-WGR),48.797 ± 0.054 (NL-WGR), and 38.716 ± 0.040 (NFI-WGR). The immediate change from the implementations of the original algorithm without guard rails is that now the LR model has the worst NRMSE with the models incorporating increased complexity subsequently improving, as expected. This agrees with the hypothesis that many data elements contain curvilinear relationships within the variable as observed with some economic indicators as well as first-order interactions. Checking to ensure that bias is not a factor with either missingness or the rate of convergences, the indications appeared weak at best. The NRMSE of the data elements were contrasted against the number of missingness within the data element producing an average correlation coefficient of 0.15. Administering a similar test against the iteration of convergence, the correlation coefficient was -0.25. If the data element did not converge, the iteration was designated as $N = 21$, which is not necessarily true and may underestimate the correlation strength. With a negative correlation coefficient, it appears that data elements that converge later may benefit from a lower NRMSE. The maximum NRMSE always came from an iteration 2 data element and many outliers disappeared after iteration 7. A future modification to the algorithm may include pausing the stopping condition check until at least seven iterations have concluded to benefit from a closer threshold in reproducing known values with the imputations.

Macro aspect

The benefit of variable range-based guard rails brought the imputations into a more plausible and defendable range of values. However, concerns of multicollinearity between independent variables used within the imputation models are still present despite the variable range-based guard rails. To dispel the concerns or minimize collinearity, collinearity discounts were applied to the variable selection process. Depending on the model, the discount had varying effects concerning convergence as illustrated in Fig. 5. The LR model benefited from increased convergence rates with

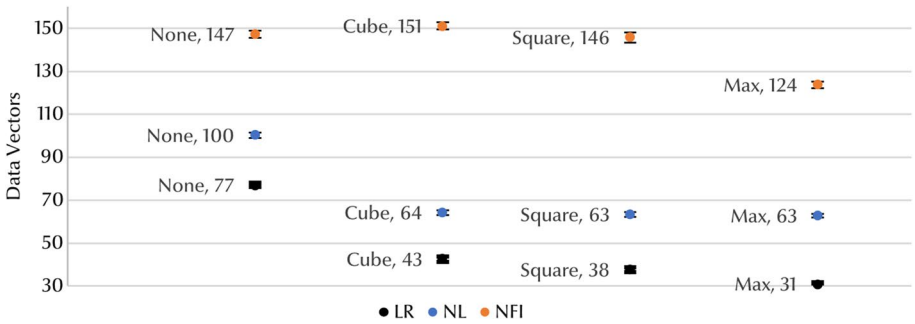

Fig. 5 Remaining unconverged data elements, iteration 20, N = 20

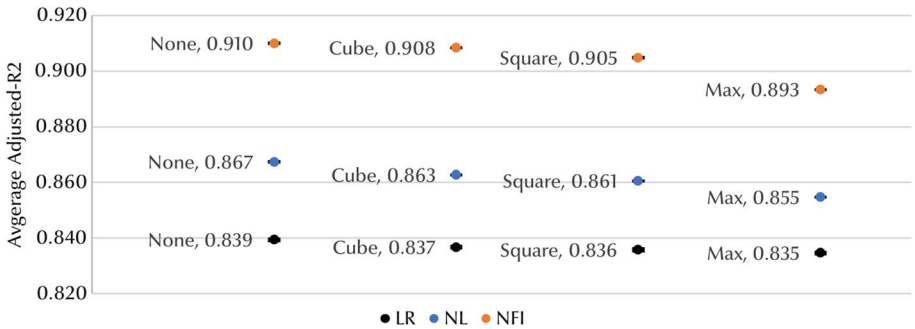

Fig. 6 Discount model average adjusted-R^2, iteration 20, N = 20

each degree of discounting. The standard error between runs was small regardless of model or discount combination averaging just over 1 data element. The NL models experienced an initial improvement toward convergence with the cubed degree of discounting, but subsequent degrees of discounting were statistically the same. NFI models saw the opposite effect. No discounting and the squared degree of discounting were statistically the same, while max discounting saw appreciable benefit in convergence. Although the discounting saw gains in convergence, it remains unclear if faster convergence produces more plausible imputations as the other metrics would indicate more defensible results from the slower NFI model rather than the faster LR model.

The average adjusted-R^2 values tells a different story. The increasing degree of discounting for all models reduced the adjusted-R^2 as seen in Fig. 6. The standard error for N = 20 was extremely tight with a maximum of 0.0003, meaning each model-discount pair were statistically different. As with the comparison between with variable range-based guard rails and without guard rails, the difference in adjusted-R^2 is small. However, with all models showing similar trends and small standard error, multicollinearity does not appear to be as big of an influence as first feared. This isn't to say that high multicollinearity does not exist, but that it does not hinder the development of plausible imputations. When the dependent variable is highly correlated to at least one dependent variable, then the R^2 value should be high. The problem with multicollinearity surfaces when multiple dependent variables are highly correlated so that the coefficients cannot differentiate stable relationships to the dependent variable.

In other words, there may be multiple solutions to the coefficients to exact the same value to the dependent variable. As mentioned previously, this is a problem of analysis, not necessarily a problem with result. Seeing how adjusted-R^2 penalizes adding independent variables of little value, high adjusted-R^2 compared between model discounts infers more defensible imputations.

Verifying the assumption that higher adjusted-R^2 leads to more defensible imputations is supported by the NRMSE metric. Lower deviation from the known value is better and the NRMSE results shown in Fig. 7 confirm that the high adjusted-R^2 of NFI produces lower NRMSE than the other models. The translation for the LR and NL models is consistent with the adjusted-R^2 results, however, the NFI is less clear. Again, the standard error is tight signifying that all model-discount pairs are statistically different. The cube and square degree of discount for the NFI model produces imputations closer to their known values over using no discount. But discounting too heavily nominates independent variables that are too far removed from alternate variables that have higher correlation values with the dependent variable. Looking at each of the validation metrics supports a different model-discount approach. However, the NRMSE defense could be weighted the heaviest by explicitly connecting imputations to known values. With the NFI model in agreement between NRMSE and adjusted-R^2 concerning the best modeling approach, it can be concluded that multicollinearity does cause a degree of problem for generating the best imputations and that a cube discount for correlation selection is warranted.

Comparative aspect

Thirty imputed complete datasets were generated for comparative testing by configuring the methodology to allow for first-order interaction while nominating variables with a cubed correlation discount following with variable range-based guard rails. Since it is computationally challenging to quickly test all data elements, this report compares only three data elements, one each from three different categories: low (< 5%) missingness with quick convergence, high (> 50%) missingness with quick convergence, and significant missingness (20–50%) without convergence. The low missingness converged on iteration 2 requiring only 1 value for imputation. The high missingness converged on iteration 6 requiring 1437 imputed values, or 83.1% of the data element. The significant missingness required 526 imputed values, or 30.4%.

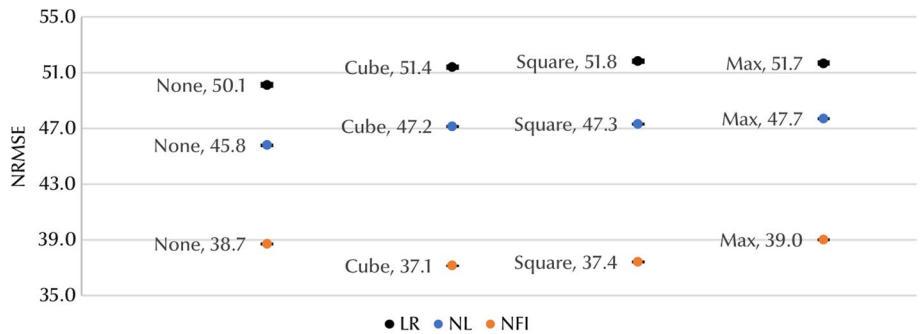

Fig. 7 Discount model NRMSE, iteration 20, N = 20

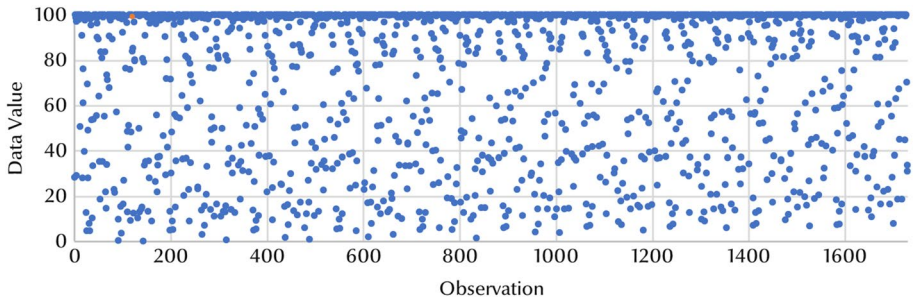

Fig. 8 Converged data element, 0.1% missingness

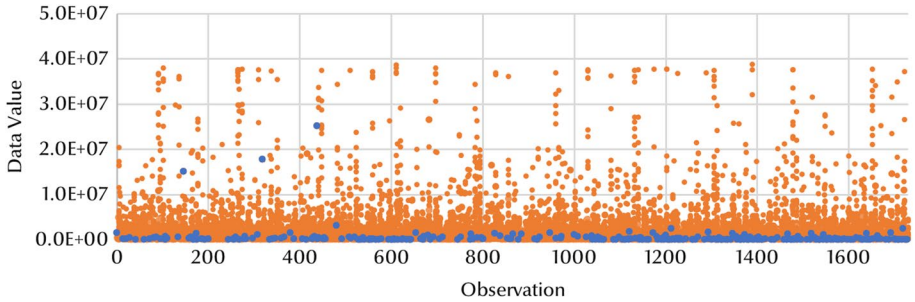

Fig. 9 Converged data element, 83% missingness

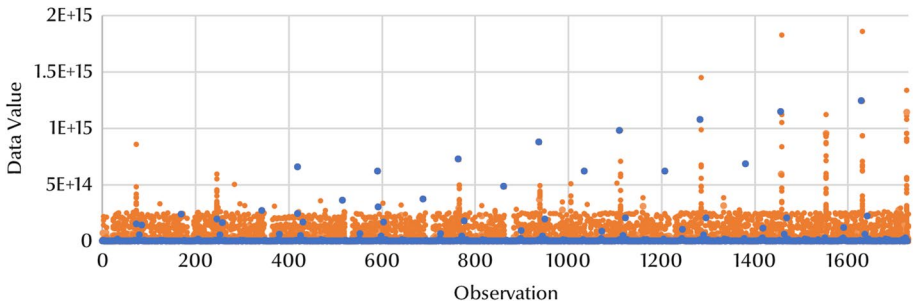

Fig. 10 Non-converged data element, 30% missingness

The scatterplots for the three datasets are provided in Figs. 8, 9, 10. The blue points indicate the known data points, while the orange points indicate the imputed data for the selected variable across all 30 imputed datasets. In Fig. 8, the one missing value in low missingness had an imputed value varying between 100 and 95.36. The missing value was in the year 2006, with the other nine years showing 100. One might assume that 2006 would also be 100, but the stochastic nature of the unknown allows for a chance of deviation. In Fig. 9, the high missingness tells a different story. Each column of orange data points shows up to 30 alternative values. At first sight, there may be questions about the plausibility, however, the descriptive statistic of standard deviation places the plausibility into perspective. For further analysis, the standard deviation across years was assessed for each country using the known data. When the same was accomplished for the imputed data, no country exceeded the maximum standard deviation of the known data, allowing the variability shown in the scatterplot. As for the non-converged data in

Fig. 10, the standard deviation analysis was not as straightforward, where all 30 datasets had a high maximum standard deviation. The maximum standard deviation of the known data was 2.33E14 whereas the imputed data ranged between 2.78E14 and 6.90E14. As a positive, it appears that the known data may see trends of increasing values over time as observations 1–173 are in year 2006 and subsequent ranges proceeding by year. The imputed values also demonstrate that potential movement. The takeaway from all three figures is that the imputed values appear to be within a reasonable distribution of the known data.

Concerning the inferential tests, it was no surprise that all three tests showed no statistical significance when comparing the distribution of the 30 generated datasets to each other in the low missingness scenario. The imputed values were all within the range of known values and it was unlikely that one data point would significantly skew the mean, median, or distribution shape. The high missingness example saw a significant difference in mean for 16 of the 30 datasets. Furthermore, all 30 datasets saw p-values below 0.05 signifying statistical differences in the median and distribution shape. Despite these results, the consideration of high missingness and the MAR assumption could still find the results plausible. The imputed values could be categorically from samples that either are adverse from measuring or are difficult to measure, as expressed in the earlier examples of the *Corruption Perception Index* or the Democratic Republic of Korea. Similar findings were observed in the non-converged example; 6 datasets demonstrating statistical differences in mean and all 30 datasets demonstrating statistical differences with the WMW and AD tests.

Summary

The original Large Dataset Imputation through Correlation-based Regression method [6] demonstrated much promise through a multiple imputation stepwise correlation approach. It provided a balance between the analyst's trade-off of time, computational power, and accuracy. Two main concerns of this original approach revolved around multicollinearity and the potential for extreme outlier values. This paper alleviates both of those concerns through exploring a full range of discounts to the variable nomination process and bounding imputation estimates within a variable ranged-based guard rail process. Both processes strengthened the plausibility and defensibility of the imputed results.

Multicollinearity is a problem of analysis in determining coefficients for cause and effect, rather than a bias in output. The None discount demonstrated superior results in the LR and NL models. Only when a small degree of discounting was applied to the NFI model did any perceived effect of collinearity surface resulting in the Cube discount being superior for the dataset considered. However, specifying the appropriate model type, from LR to NL to NFI, demonstrated greater gains than the effects of discounting collinearity.

To further enhance the prior approach, variable range-based guard rails were developed that bounded the imputations into a plausible range and deterred subsequent iterations within the algorithm to exacerbate outliers. In hindsight, it aligns with the superiority of pmm on small datasets where imputations are likewise bounded to values

already seen in the dataset. Unlike pmm, the variable ranged-based guard rails allow values that are probable in the distribution yet not observed, widening the aperture for plausible values.

Providing three aspects of analysis assisted in quantifying progress while increasing the defensibility of the method. The micro aspect analysis highlighted the improvements in convergence rates of individual data elements compared to [6] while maintaining strong goodness of fit. The macro aspect analysis quantified how little collinearity effects hinder the imputation through the adjusted-R^2 results demonstrating decreasing values with discounting and all but the interactions modeling showing lower NRMSE without discounting, dispelling concerns over using a correlation-based selection process. The comparative aspect analysis visualized the imputations to the known value distributions for a qualitative approach to plausibility. The inferential tests conducted alongside the visual assessment and descriptive statistics demonstrated opposing theories on plausibility, which cautions analysts from relying on a single metric when evaluating imputations. When working with MAR and NMAR data, an expert in the data is necessary for more conclusive analysis.

Outside of overcoming numerical problems in generating imputations, the improved approach also provided insight into the rate of convergence. Rather than providing a user-specified static tolerance for a stopping condition, the approach relied on the data itself to generate tailored data element tolerances by exploiting the residuals in modeling the known data. The concept leans on the definition of stochastic convergence of the r-th order mean where the difference of successive iterations is statistically zero. Using the distribution of the residuals captured in the first iteration of only known values, which were also used for noise, the algorithm conducts a check between iteration N and $N+1$ to measure the difference between estimates. Should the difference be within 3 standard deviations of the distribution of residuals convergence is assumed and the stopping condition applied.

Although the MASS-impute algorithm improved the original correlation-based approach, there are still areas that require further refinement. It was noted that the worst NRMSE values were captured during the first iterations, so further modifications to the algorithm may investigate not allowing stopping conditions until after a set number of iterations. Such changes would increase the processing time of the algorithm, but at the potential benefit of improved accuracy. The investigation would illuminate the trade space between these two analytical trade-offs for balancing out the algorithm's parameters. Additionally, as seen in the comparative analysis, some of the high standard deviations in the imputations continue to be a concern. Known outliers in some datasets may be allowing too much variability in the noise element of the algorithm. The high missingness scatterplot showed three known values that would pull at the regression line used to generate the noise residuals. These outliers could potentially be adding too much variability to the stochastic nature of the estimates, especially when the outliers are more prevalent as in the non-converged example. Future modifications may investigate better accounting for these outliers when producing the pool of noise.

Using MASS-impute, the multiple imputations appear plausible while dispelling concerns about variable selection based on correlation. As with the finding in [6], the evolution of the methodology continues to balance computation time, power and accuracy in

achieving traceable, defensible imputations for large datasets, including those that may exhibit over 20% missingness for some variables.

Abbreviations

AD	Two-sample Anderson–Darling test
FNI	Nonlinear with first-order interactions
LR	Linear
MAR	Missing at random
MASS-impute	Multicollinearity applied stepwise stochastic imputation
MCAR	Missing completely at random
MICE	Multiple imputation by chain equations
NL	Nonlinear
NMAR	Missing not at random
NRMSE	Normalized root mean square error
Orig	Original
UN	United Nations
WGR	With variable range-based guard rails
WMW	Wilcoxon–Mann–Whitney test

Acknowledgements
Not applicable.

Author contributions
Both authors read and approved the final manuscript.

Funding
No funding received.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This article does not contain any studies with human participants or animals performed by any authors.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 25 July 2022 Accepted: 29 January 2023
Published online: 15 February 2023

References

1. Chhabra G, Vashisht V, Ranjan J. A comparison of multiple imputation methods for data with missing values. Indian J Sci Technol. 2017;10(19):1–7. https://doi.org/10.17485/ijst/2017/v10i19/110646.
2. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst. 2012;32(1):77–108. https://doi.org/10.1007/s10115-011-0424-2.
3. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Appl Artif Intell. 2019;33(10):913–33. https://doi.org/10.1080/08839514.2019.1637138.
4. Ahner D, Brantley L. Finding the fuel of the Arab spring fire: a historical data analysis. J Def Anal Logist. 2018;2(2):58–68. https://doi.org/10.1108/JDAL-03-2018-0008.
5. Kane ZJ. An imputation approach to developing alternative futures of country conflict. Master's thesis, Air Force Institute of Technology; 2019.
6. Leiby BD, Ahner DK. A large dataset imputation approach applied to country conflict prediction data. Int J Math Comput Sci. 2022;16(3):11–7.
7. Wayman JC. Multiple imputation for missing data: what is it and how can i use it? In: Annual meeting of the American educational research association, Chicago, vol. 2; 2003. p. 16. https://doi.org/10.1002/0471264385.wei0204.
8. Lodder P. To impute or not impute : that's the question. In: Mellenbergh GJ, Adér HJ, editors. Advising on research methods: selected topics (2013). Huizen: Johannes van Kessel Publishing; 2013. p. 1–7.
9. Arel-Bundock V, Pelc KJ. When can multiple imputation improve regression estimates? Political Anal. 2018;26(2):240–5. https://doi.org/10.1017/pan.2017.43.
10. Plumpton CO, Morris T, Hughes DA, White IR. Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. BMC Res Notes. 2016;9(1):1–16. https://doi.org/10.1186/s13104-016-1853-5.

11. Lindner T, Puck J, Verbeke A. Misconceptions about multicollinearity in international business research: identification, consequences, and remedies. J Int Bus Stud. 2020;51(3):283–98. https://doi.org/10.1057/s41267-019-00257-1.
12. Nguyen CD, Carlin JB, Lee KJ. Practical strategies for handling breakdown of multiple imputation procedures. Emerg Themes Epidemiol. 2021;18(1):1–8. https://doi.org/10.1186/s12982-021-00095-3.
13. Disatnik D, Sivan L. The multicollinearity illusion in moderated regression analysis. Mark Lett. 2016;27(2):403–8. https://doi.org/10.1007/s11002-014-9339-5.
14. van Buuren S, Groothuis-Oudshoorn K. Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
15. Harvey AC. Miscellanea: some comments on multicollinearity in regression. Appl Stat. 1977;26(2):188–91.
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(85):2825–30.
17. Kearney J, Barkat S. Autoimpute documentation; 2021. https://readthedocs.org/projects/autoimpute/downloads/pdf/latest/.
18. van Buuren S. Flexible imputation of missing data. 2nd ed. Boca Raton: CRC Press; 2018.
19. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18(1):50–60. https://doi.org/10.1214/aoms/1177730491.
20. Jäntschi L, Bolboacă SD. Computation of probability associated with Anderson–Darling statistic. Mathematics. 2018;6(88):1–16. https://doi.org/10.3390/math6060088.
21. Engmann S, Cousineau D. Comparing distributions: the two-sample Anderson–Aarling test as an alternative to the Kolmogorov–Smirnoff test. J Appl Quant Methods. 2011;6(3):1–17.
22. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. SciPy 1.0 contributors: SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.

Publisher's Note