What's missing? The effect of missing data and imputation techniques on predictive performance in forecasting civil war violence

David Randahl

Department of Peace and Conflict Research, Uppsala University, Email: david.randahl@pcr.uu.se

Abstract

Missing data is a ubiquitous problem in science in general, and social science in particular. It is well-known that not dealing properly with the missing data problem may cause inferences to be biased. Less attention has been given to the implication of missing data and imputation techniques on the predictive performance of different models. This paper explores these issues through generating missing data in a data set used for forecasting the number of fatalities from political violence in all countries during 1989-2021. The results show that single model-based and non model-based imputation techniques perform best with regards to the predictive performance of the model on the outcome of interest, while multiple imputation techniques fare generally fare worse. These results highlight that the guidelines for how to handle missing data in inferential studies are not generally transferable to forecasting studies. Instead, this paper suggests that the imputation technique should be specifically tailored to the research objective and to the variables which are to be imputed.

1 Introduction

Missing data is a common problem in almost all fields of science, but especially so in fields which rely on human collected data. Missing data refers to when some observations in a data set have incomplete data i.e. when information is missing for some, but not all, variables. It is well-known that the method for handling the missing data affects inference, as parameter estimates may be biased and associated uncertainty metrics may be either too small or too large if an inappropriate method for handling missing data is used (see for instance Honaker et al., 2010; van Buuren, 2018; Lall, 2016). The effects of different missing data techniques on predictive performance of models are, however, less studied, with only a few systematic studies showing what the effects of missing data are on predictive performance, and how different imputation techniques affect this performance (exceptions include Poulos and Valle, 2018; Batista and Monard, 2003; Twala, 2009).[1]

The lack of knowledge about the effects of missing data treatment for predictive studies is especially problematic as existing imputation techniques are developed with the aim of producing unbiased inferences with correct uncertainty metrics, rather than with the goal of producing accurate predictions of some outcome of interest (Hollenbach et al., 2021; van Buuren, 2018; Honaker, King, and Blackwell, 2011). Compounding this issue is that these existing imputation techniques are primarily designed with parameter estimates for GLMs in mind, and may thus have additional adverse effects on forecasting models such as tree-based models. In packages for supervised learning such as `caret` for `R` and `scikit-learn` for Python, imputation is often just briefly discussed as a necessary step of data pre-processing, with relatively little or no discussion of how the choice of imputation technique affects the performance of the model (Telenczuk, 2022; Kuhn, 2019; Pedregosa et al., 2021). In all, these deficiencies leave forecasting oriented researchers mostly in the dark with regards to questions of how to best handle missing data and the what effects of the choice of imputation technique. As supervised learning and forecasting oriented research are becoming increasingly common and attractive in a range of fields within social science as well as other fields of science (see for

[1] There also exist a number of studies which compare different imputation techniques for forecasting studies but evaluate the imputation techniques in terms of the accuracy of the imputations rather than the effect of the imputation techniques on predictive performance (see for instance Jadhav, Pramod, and Ramanathan, 2019; Khan and Hoque, 2020; Tang and Ishwaran, 2017). This is, however, not necessarily an appropriate evaluation metric for imputations as the goal of imputation is not prediction. This point is further expanded on in section 2.

instance Ward, Greenhill, and Bakke, 2010; Hegre et al., 2019; Muchlinski et al., 2016; Montgomery et al., 2015; Schutte et al., 2021) these questions are getting ever more important.

This paper aims to improve the knowledge of the effects of missing data techniques on predictive performance of forecasting models, and to offer some initial guidance to forecasting researchers on how the choice of imputation technique affect the predictive performance of their models. To achieve this, this paper uses a complete panel data set used for forecasting civil war violence (Hegre et al., 2019) and in this data set simulates missing data under different assumptions. The missing data are then imputed using a range of different single- and multiple imputation techniques after which two forecasting models, one random forest model and one linear regression model, are trained on the imputed data. The out-of-sample predictive performance of the forecasting models trained on the imputed data are then compared to the performance of the models trained on the full, complete, data when making predictions on a held out test data set. The results show that the choice of imputation technique does affect the predictive performance of the tested models. This effect is greatest for the random forest regressor which is substantially hurt by missing data and inappropriate imputation strategies. Imputation using the global mean fare poorly across all tests, while group-wise mean imputation, last observation carried forward/backward, single imputation based on random forests, and multiple imputation based on a two-level model perform yield the best predictive performance. Contrary to the case in inferential studies, the single imputation techniques tested generally yield better performance than their multiple imputation counterparts. The only exception to this is the two-level normal model which yields good performance, but at a high computational cost.

The remainder of the paper is organized in four sections. The first section reviews the literature on missing data mechanisms and existing imputation techniques. This is followed by a section describing the data and prediction problem, as well as the setup for the simulation study. In the third section, the results of the simulation study are presented and discussed, and in the final section conclusions are drawn.

2 Missing data mechanisms and methods for handling missing data

Missing data is usually divided into three different categories depending on the mechanism which cause the data to be missing. These categories are data which are *Missing Completely at Random* (MCAR), data which are *Missing at Random* (MAR), and data which are *missing not at random* (MNAR). Using the notation of van Buuren (2018), let Y be an nxp matrix containing some data on p variables for n units, and R be a nxp response indicator where the individual elements are denoted as $y_{i¦j}$ and $r_{i¦j}$ respectively. Further, let $r_{i¦j=,}$ if $y_{i¦j}$ is observed and $r_{i¦j}\ = \bullet$ if $y_{i¦j}$ is missing and Y_{obs} be all observed elements of Y and Y_{mis} be all missing elements of Y, and   be the parameters of the missing data model. When data are MCAR then the probability of an individual value in Y being missing is independent both of the missing value itself, as well as all observed information for the case. Formally we can express this as

$$\mathrm{Pr}^1 R\ =\ \bullet \mathrm{j} Y_{obs¦} Y_{mis¦}\ \ ^{\mathrm{o}} = \mathrm{Pr}^1 R\ =\ \bullet \mathrm{j}\ \ ^{\mathrm{o}}$$

This implies that there is no information in Y which can help us predict whether or not an individual value is missing or not. Data may for instance end up being MCAR if there are random coding errors by human coders, if certain swathes of data are lost in an accident, or if respondents to a survey would with equal probability decide whether or not to answer a specific question. If, on the other hand, there is other information in Y outside the specific variable where the data is missing which can help predict whether or not the specific value is missing or not, then the data is considered to be MAR. Formally this implies that

$$\mathrm{Pr}^1 R\ =\ \bullet \mathrm{j} Y_{obs¦} Y_{mis¦}\ \ ^{\mathrm{o}} = \mathrm{Pr}^1 R\ =\ \bullet \mathrm{j} Y_{obs¦}\ \ ^{\mathrm{o}}$$

If, for instance, countries with a lower level of GDP/capita are less likely to report the education level in the country, or if survey respondents belonging to the Republican party are less inclined to answer survey questions about their income, then these data are MAR. Finally, if the likelihood of a specific value being missing is dependent on the value itself, then the data is MNAR. Formally this implies that the expression

$$\Pr(R = 0|Y_{obs}, Y_{mis})$$

does not simplify. This would, for instance, refer to a situation where countries with a lower level of GDP/capita is less likely to report the figure of GDP/capita, or where survey respondents with higher incomes are less likely to respond to a question of their income (King et al., 2001; Lall, 2016).

2.1 Methods for handling missing data

To handle missing data researchers generally either use listwise-deletion, in which all observations which contain at least one missing value are removed from the data set, or imputation, in which the missing values are replaced with some other (plausible) value. Within imputation there exists a wide range of different options, from simple non model-based methods such as imputation using the mean or mode value of the specific variable or last observation carried forward/backward (LOCFB) in panel data, to imputation techniques based on predictive models such as GLM:s, CART, or random forests (see for instance Honaker, King, and Blackwell, 2011; Batista and Monard, 2003; van Buuren and Groothuis-Oudshoorn, 2011; Shah et al., 2014). Imputation are further divided into single imputation methods, where the missing values are replaced once, and multiple imputation methods, where the missing values are replaced multiple, *m*, times and the analysis conducted on each of the imputed data sets after which the results are pooled (van Buuren, 2018; King et al., 2001).

Which method(s) for handling missing data that are appropriate is dependent on the mechanism which cause the data to be missing and where the missingness occurs. In inferential research, simple non model-based methods such as mean or median imputation and LOCF are almost never appropriate as they may cause parameter estimates to be biased even under MCAR, and may cause standard errors to be too small. Single imputation methods based on predictive modelling may produce unbiased estimates but instead cause the standard errors to either be too small or too large (Graham, 2009; van Buuren, 2018). Listwise-deletion produce unbiased, albeit less efficient, parameter estimates if the data are MCAR or have missingness only in the independent variables of a GLM. The gold standard of missing data techniques are multiple imputation techniques which produce unbiased parameter estimates which are also more efficient compared to estimates from listwise deletion under both MCAR and MAR. When data are MNAR, on the other hand, both listwise-deletion and multiple imputation produce biased parameter estimates (Pepinsky, 2018; van Buuren, 2018).

An important aspect to keep in mind when discussing imputation methods for inferential research is that imputation *is not* prediction, i.e. the goal of imputing data is not to replace the missing values with values as close as possible to the true values. Rather, the goal of imputation is to replace the missing values with values which cause the resulting parameter estimates to be unbiased and have appropriate measures of uncertainty, i.e. so that their standard errors enter the Goldilocks range of being neither too large or too small but just right. This means that the imputation model which produce the most accurate values in the imputed data set is not necessarily the best imputation method, as these methods may bias the standard errors of the resulting analysis (van Buuren, 2018).

2.2 Missing data in forecasting versus inferential research

That missing data mechanisms and methods for handling missing data has almost exclusively been discussed in terms of inferential research poses a problem for researchers interested in forecasting, as the concepts which are important in inferential research, whether parameter estimates are unbiased and have appropriate standard errors, are not necessarily as important in forecasting. Instead, forecasting oriented researchers should primarily be interested in what the effects of missing data and imputation are on the predictive performance of their models (Shmueli, 2010). This means that the supremacy of multiple imputation methods over single imputation methods with higher accuracy cannot necessarily be guaranteed. Similarly, the severity of the missing data problem cannot necessarily be ranked in terms of whether the data is MCAR, MAR, or MNAR. For instance, if a researcher is interested in classification of an outcome and the data are MNAR on the outcome, then the MNAR mechanism may operate similar to a downsampling routine if missingness is more common in the dominant class. In this case, the predictive performance of the classifier may *increase* as a result of the missingness, since downsampling routines are known to increase the predictive performance of certain classifiers (Japkowicz and Stephen, 2002; Muchlinski et al., 2016).

An additional complicating factor is that the standard missing data treatment in most studies, listwise deletion (Lall, 2016), makes it impossible to make forecasts for the observations with missing values. If the missingness occurs in the test data, on which the forecaster aims to evaluate the model, then the data has to be imputed in some way in order to make valid predictions. This also means that the values which are imputed need to be within the range of plausible values for the variable as the resulting predictions may otherwise end up being unreasonable (Graham, 2009; Allison, 2001).

In all, there are many unanswered questions about what the effects of missing data and the methods for handling missing data are on the predictive performance of forecasting models. This paper will focus specifically on a subset of five of these questions, namely:

(1) What are the effects of different methods for handling missing data under different missingness mechanisms on predictive performance of models?

(2) Does multiple imputation perform better than single imputation with regards to predictive performance?

(3) What is the effect of the number of multiply imputed data sets, m, on the predictive performance of models?

(4) How does the accuracy of the imputations affect the predictive performance of models?

(5) Do the answers to 1-4 differ depending on the type of forecasting model used?

This paper aims to shed some initial light on these questions by simulating missingness and imputing data in a complete data set on forecasting civil war violence. The experimental setup of the study is outlined in the section below.

3 Experimental setup and forecasting problem

To test the questions outlined in Section 2 above, this paper will use a forecasting model from the Violence Early Warning System (ViEWS) forecasting system (Hegre et al., 2019) and introduce simulated missing data into the data. The data consists of a panel data set of country-month observations for all countries in the time period 1989-2021, with the forecasting target being the natural logarithm of fatalities from state-based violence[2] one month into the future.[3] The specific forecasting model contains 23 predictor features, including features derived from newspaper topic models (Mueller and Rauh, 2018), economic and demographic features from the world development indicators (World Bank, 2021), and political indicators from the Varieties of Democracy project (Coppedge et al., 2021). All predictors and the forecast target are continuous variables and descriptive statistics and densities for the variables can be found in Table 3 and Figure 6 in the appendix.

The full data consists of 62,310 complete country-month observations across 207 countries. To ensure proper evaluation of predictive performance of the forecasting models, the data was further divided into a training data set, containing all data up until 2016, and a test data set containing data in the period 2017-2021 on which the predictive performance was evaluated on. This left 55,740 complete observations in the training partition and 6,570 complete observations in the test partition.

Missing data are then simulated in the complete data using multivariate amputation which allows for proper missing data generation under different missingness mechanisms and conditions. Multivariate

[2] $\log_{,p}$, as the un-transformed target contains zero-values.

[3] For a more detailed description of the forecasting target, see Hegre and ViEWS Team (2021)

amputation can be thought of as imputation in reverse and generates missingness based on weighted sum scores, equivalent to linear regression estimates, transformed to a logistic distribution (Schouten, Lugtig, and Vink, 2018). For this experiment data are simulated to be missing on a grid of conditions, depending on the missingness mechanism, MCAR, MAR, and MNAR, the proportion of observations with missing values, 1%, 5%, 10%, and 20%, whether the missingness is restricted to the predictor features, or to both the predictor features and the target, and whether the missingness occurs in the training partition, test partition, or both. In total, this leads to 72 different simulation setups for the missingness, summarized in Table 1. In order to avoid any potential leakage between the partitions, the data are amputed separately for the training and test partition.

Table 1. Experimental settings used in the simulations

Experimental setting	Values
Missingness mechanism	MCAR, MAR, MNAR
Missingness level	1%, 5%, 10%, 20%
Missing variables	Predictors, predictors and target
Missing in partition	Training, test, both
Number of repeats	100
Total number of simulations: 7,200	

After amputation the data are treated with a wide range of missing data techniques, from simple solutions such as listwise deletion and mean imputation, to single model-based imputation techniques using the R package *VIM* (Kowarik and Templ, 2016) and multiple imputation techniques using the R packages *Amelia* (Honaker, King, and Blackwell, 2011), *mice* (van Buuren and Groothuis-Oudshoorn, 2011), and *miceadds* (Robitzsch and Grund, 2021). A complete list of the 13 missing data treatment techniques can be seen in Table 2. These 13 missing data treatment techniques are not supposed to be seen as an exhaustive list of possible techniques for handling missing data, but rather as a subset of commonly used missing data techniques which vary across the across the non model-based, single, and multiple imputation spectrum and across different levels of complexity and computational intensity among the model-based imputation strategies. All model-based imputation techniques use the same variables which are allowed to be missing (i.e. all predictor features or all predictor features and the prediction target) as predictors in the imputation model and use standard settings for optional arguments.

To avoid potential leakage between the partitions (see for instance Kapoor and Narayanan, 2021), the training and test partitions are imputed separately. Listwise-deletion is only applied for the setup where data is missing in the training partition, as listwise deletion cannot generate complete predictions when missingness occurs in the test partition. All multiple imputation techniques included are set to impute a total of $m = f\bullet$ data sets.

After each missing data technique has been applied, two different forecasting models are trained on the training partition; one random forest model and one linear regression model using all predictor features in the data set. For data which have been multiply imputed, training of the models takes place once for each imputed data set. Predictions for the multiply imputed data sets are treated as naive ensembles, i.e. then taken to be the simple average of predictions across all imputed data sets.[4]

[4] When data are only missing in the training partition or the test partition, this is the average across m predictions, when data is missing in both the training and test partitions, m multiply imputed data sets are generated in both the training and test partition, leading to an average of m^f predictions per observation.

Table 2. Missing data handling techniques evaluated

Missing data technique (abbreviation)	Implementation in R
non model-based solutions	
List-wise deletion (lwd, only missing in training)	
Global mean imputation (mean)	mice::mean
Group-wise mean imputation (2lmean)	mice::2lmean
Last observation carried forward/backward (locfb)	
Single model-based imputation solutions	
Random forest imputation (ranger)	VIM::rangerImpute
Iterated model-based imputation (irmi)	VIM::irmi
Regression imputation (reg)	VIM::regressionImp
k Nearest Neighbor imputation (knn)	VIM::kNN
Multiple imputation without panel structure	
Predictive mean matching (pmm)	mice::pmm
Bayesian linear regression (norm)	mice::norm
Expectation-Maximization (EM) Bootstrapping (amelia)	amelia::amelia
Multiple imputation with panel structure	
Two-level predictive mean matching (2lpmm)	miceadds::2l.pmm
Two-level normal model (2lnorm)	mice::2l.norm

The effect of the missing data treatment on the predictive performance of the model is then calculated by making predictions from the models trained on the treated training partition on the treated test partition and calculating two quantities based on the difference in the mean squared error (MSE) between the model trained on the complete data and the model trained on the imputed data.[5] The first of these is the difference in MSE for all observations in the test data, and the second is the difference in MSE for all observations which have at least one missing value in the test partition. The MSE on the test partition for the models trained on the full training data are 0.726 and 1.542 for the random forest and linear regression models respectively. To evaluate question (3) from section 2, the effect of m on predictive performance, a subset of the multiple imputation techniques were tested on all m from , to ,•• and evaluated across all m. To evaluate question (4), relating to the effect of the accuracy of imputations on the predictive performance, the normalized root mean squared error (NRMSE) (Stekhoven and Bühlmann, 2012) for all missing observations is calculated on the training partition for each imputation technique used. NRMSE is used to measure the accuracy of the imputations rather than MSE since the imputed variables are all measured on different scales. For multiply imputed data sets, the NRMSE is calculated as the mean NRMSE across all m imputed data sets.

4 Results

Figure 1 below shows the effect of using listwise deletion on missing data in the training partition on the MSE in the test partition for the two different models. These results show some interesting patterns. First, it should be noted that the effect of missing data is generally larger and more uniform for the random forest where the MSE consistently rises with the proportion and severity of missing values. Second, for the linear regression model it seems clear that data which are MCAR do not seem to have any systematic effect on predictive performance. Additionally, in the linear regression model, certain combinations of missingness

[5] The prediction target is never imputed in the test partition and all evaluation are against the true target value

mechanisms and location of the missing data seem to produce increased predictive performance, especially when data are solely missing in the predictor data and when data are MAR or NMAR. This result is likely related to which cases are missing under the different missingness simulations. The ampute function generating the missing values are set to ampute values in the right side of the distribution, and if this would cause a re-balancing of the training data towards more representative cases when data are missing only in the predictor features but towards less representative cases when data are missing in both predictors and target, then such heterogeneous effects could appear. Amputing the data on a different part of the distribution could therefore possibly cause the missing data to have different effects on the predictive performance. Overall, the predictive performance is worst affected by using listwise deletion when data are MAR or NMAR in both the predictor data and the target.

Figure 1. The effect of using listwise deletion on missing data in the training partition under different simulations of missingness. Point estimates are medians across all 100 simulations, and intervals span the 10th-90th percentiles of values. Note: the proportion of missing observations are a binned scale with allowed values at 1%, 5%, 10% and 20%. Positive values indicate worse performance

Figure 1 clearly shows that treating data in the training partition with listwise deletion is not unproblematic with regards to the predictive performance of models. This is especially true for the random forest model, and highlights the importance of imputation to treat this problem. The occasional positive effect of listwise deletion in the linear regression case is likely due to an effect similar to downsampling, where the exclusion of certain cases increases the overall performance of the model. In this case the data should reasonably be downsampled or re-weighted using a researcher specified downsampling routine, rather than through listwise deletion of missing data.

Figures 2-3 below show the impact of the tested imputation methods on the predictive performance of the linear regression and random forest models across the different simulations of missingness for the two different evaluation metrics. The full results for all models are found in the appendix. These results show that all imputation techniques, except mean imputation, help alleviate the problem of missing data in prediction.

Figure 2. The effect of different imputation techniques on the overall predictive performance of the linear regression and random forest models under different simulations of missing data in both the training and test partitions. Point estimates are medians across all 100 simulations, and intervals span the 10th-90th percentiles of values. Multiple imputation methods measured form=20.

Figure 3. The effect of different imputation techniques on predictive performance for observations with at least one missing value in the test partition of the linear regression and random forest models under different simulations of missing data in both the training and test partitions. Point estimates are medians across all 100 simulations, and intervals span the 10th-90th percentiles of values. Multiple imputation methods measured for m=20.

The results also show that non model-based and single imputation techniques generally perform best with regards to both the effect on the overall predictive performance of the model (Figure 2) and the predictive performance on the observations with at least one missing value (Figure 3). Overall there are relatively small differences between the best performing imputation techniques, the group-wise mean imputation, last observation carried forward/backward, single k nearest neighbor imputation, single random forest imputation, and the two level normal model. Across these it is worth noting that the last observation carried forward/backward and k nearest neighbor imputation methods seem more robust to data which are NMAR with missingness in both the predictor and target features. It is also worth noting that the regression and iterated regression methods, as well as the multiple imputation methods apart from the two-level normal model, fare substantially worse than the single imputation methods (both model-based and non model-based).

In line with expectations, the imputation techniques fare substantially worse compared to the true/full model when the data are MNAR and missing among predictor features and the prediction target. This is true for almost all imputation techniques, except for the LOCFB and knn imputation techniques when using a random forest model where the performance is roughly on par for this type of missingness as for other types of missingness. Somewhat surprisingly a few of the imputation techniques produce negative MSE differences, indicating a positive effect on the predictive performance, for the linear regression model. This seems to specifically be true for for the regression based single and multiple imputation techniques when the data are MAR and missing among both predictor features and the forecasting target. While this result is somewhat surprising, it may be related to the imputation models using information from the forecasting target to impute the missing values among the predictors and thereby causing an increased performance. This is would be similar to using the missing information as predictive information.

4.1 The effect of the number of imputations on predictive performance

To investigate the impact of the number of imputations on the predictive performance of the three multiple imputation techniques without panel structure assessed all m from 1 to 100 for data which was missing in the training partition.[6] The results are shown in Figure 4 below.

These results show a clear effect of increasing m above 1, i.e. using a single imputed data set from a multiple imputation technique does not yield satisfactory results in terms of predictive performance. For the linear regression model this effect is small, possibly reflecting the overall poor performance of the linear regression model on this prediction problem, and seems to reach its maximum effect around an m of 5 or 10. For the random forest model, the increase in model performance is quite steep up until an m of between 10 and 20 after which the effect of increasing m further tapers off. It should, however, be noted that the performance of the random forest model continues to increase up to an m of 100. In all, these results seem to suggest that the just as with inferential studies that a higher m is better but that a low m of perhaps 10 or 20 is sufficient to achieve a boost in performance. The m of 5 which is generally recommended in inferential studies is, however, perhaps on the lower side for optimal predictive performance (van Buuren, 2018). In the end, the optimal number of m should be set based on the trade-off between the increase in predictive performance and the computational cost of increasing m.

[6] Due to the computationally intensive nature of the multiple imputation techniques with panel structure, see Figure 7, this test was only done for the multiple imputation techniques without the panel structure and for data which were only missing in the training partition as the number of predictions generated per model would be m^f.

Figure 4. The effect of the number of imputations, M, for multiple imputation techniques on the predictive performance of the linear regression and random forest models. The effect is the mean effect across all simulation setups where missingness occurred in the training partition.

4.2 Imputation accuracy and predictive performance

Another question this paper set out to answer was the relationship between imputation accuracy and predictive performance, with predictive accuracy being calculated as the normalized root mean squared error (NRMSE) (Stekhoven and Bühlmann, 2012). The relationship between the median NRMSE and the median difference in the MSE of the post-imputation estimated model and true model can be seen in Figure 5 below, where the accuracy and predictive performance of the imputation techniques are summarized across all simulations.

These results show a moderate positive correlation between the median NRMSE and median difference in MSE for all missingness types. This positive correlation is most when evaluating the performance for the observations which have at least one missing value for the random forest model. This is in line with expectations as we would expect values closer to the true values to generate better predictions for those specific observations. In general, these results indicate that imputation accuracy indeed does have an impact on the predictive performance when imputing missing data, however, imputation accuracy is evidently not the only factor which influences whether a specific imputation technique is appropriate for the prediction problem. Rather a number of different factors such as the suspected missingness mechanism, the proportion of missing values, and whether data are casewise or intermittently missing, are, in addition to imputation accuracy, important when selecting the imputation technique for the specific imputation problem.

Figure 5. The effect of imputation accuracy on predictive performance of linear regression and random forest models under different simulations of missingness. Linear trend-line imposed in black. Mean imputation not shown to improve clarity. Point estimates are medians across all 100 simulations.

4.3 Computational considerations

The sections above have outlined the results with regards to the effect of missing data techniques on the predictive performance of different models. However, imputation is also a potentially computationally intense operation and the different imputation techniques use vastly different level of computational resources. Figure 7 in the appendix show the computational time used for imputation of both the training partition and test partition for each of the 11 model-based imputation techniques[7]. For the multiple imputation techniques the time refers to the time it takes to impute all 20 data sets.

These results show that the imputation techniques based on linear regression, reg and amelia imputations, are substantially less computationally intensive than the other imputation techniques. In the middle range of computationally intensive imputation techniques we find the pmm and norm multiple imputation techniques from the mice package, and the irmi and knn single imputation techniques from the VIM package, and in the most computationally intensive category we have the two-level multiple imputation models from the mice package (2lpmm and 2lnorm) as well as the random forest single imputation from VIM. If imputation of the data need to be repeated many times, for instance if the forecasting effort is near-real time and/or the data is large there may be substantial computational costs involved with the imputation strategy if one of the more computationally intense methods are used. Perhaps especially problematic is that the overall best performing imputation technique, the two-level normal model, is also the computationally most intense imputation technique.

Yet another aspect to keep in mind for the computational considerations of imputation is the training time involved with training machine learning models on multiply imputed data. As the multiple imputation

[7] Computations were carried out on the UPPMAX/Rackham HPC cluster where each compute node consists of two (2) Intel Xeon E5 2630 v4 at 2.20 GHz/core (10 cores, 20 threads, 25 MB LLC, and a bandwidth of 68.3 GB/s) and 128GB of ECC 2400MHz DIMM DRAM memory. All simulations were run on two cores with access to 12.8 GB RAM memory. Only the ranger/random forest imputation model was able to utilize both cores, all other imputations techniques only utilized one core. The non model-based imputation techniques are not included in the comparison as imputation from these techniques are instant or near instant.

techniques require m models to be trained, one for each multiply imputed data set, the training time for the models increase linearly with m. For small to medium data sets and a reasonable sized m this should not have too much of an impact. However, on models which are highly computationally intensive increasing m could substantially increase the computational burden of the forecaster. As shown in section 4.1 above, using an m of 10 or 20 generally seems to yield acceptable results and increasing m further may therefore not be necessary if the computational costs are deemed to be too large.

5 Discussion

The results from the simulation study above highlight a number of important results which should guide forecasting practitioners when choosing how to handle missing data. The most important results is that the guidelines used for handling missing data in the inferential case do not necessarily seem to hold up well for forecasting oriented studies. While simple solutions such as listwise deletion and global mean imputation perform poorly just as in inferential research, the best performing methods with regards to predictive performance seem generally to be single imputation methods, either non model-based such as group-wise mean imputation or last observation carried forward/backward, or model-based solutions such as k nearest neighbor imputation or random forest imputation. The only multiple imputation technique which yielded good performance was the two-level normal model, albeit at a high computational cost. While these results are based on analysis of only one data set and a small subset of potential imputation techniques, and should therefore be considered preliminary, they do suggest that the guidelines for how to handle missing data in inferential studies is not necessarily transferable to forecasting oriented studies.

If the only goal of the forecasting effort is the predictive performance of the model(s) practitioners should therefore probably use single imputation methods such as the k-nearest neighbor imputation technique, or, if the data is intermittently missing in a panel data setup, group-level mean or last observation carried forward/backward. The usefulness of these two non model-based imputation techniques are, however, likely dependent on the characteristics of the imputed variable and how much it varies over time. For instance, group-level mean is likely not a good imputation strategy for variables which tend to trend in a specific direction for cases over time. Similarly, last observation carried forward/backward is likely not an appropriate imputation method for variable which vary a lot across short periods of time. It is also important to keep in mind that these two non model-based imputation techniques only work in cases where the data is panel data or time-series data. The higher computational cost of the random forest based imputation techniques do not seem warranted given that the performance of this imputation technique is similar to the k nearest neighbor imputation technique.

The results presented in this paper are most problematic for researchers who are interested in both producing accurate forecasts and make valid inferences. It is well known that the single imputation techniques which fared best with regards to their effect on the predictive performance of the model also fare poorly if the goal is to produce regression estimates with appropriate uncertainty metrics. One option here is to use multiple imputation with the two-level normal model which performed well on predictive performance. However, this technique was also very computationally intensive and may therefore be problematic for large data sets. Another alternative in this situation would be to use different imputation techniques for the model when forecasting is the goal and when inference is the goal. While this may sound somewhat unorthodox, this should not pose any problems for the validity of the study since the goal of multiple imputation in inferential research is not to produce correct imputations, but rather to produce imputations which generate correct inferential estimates. Similarly, the goal of imputation in the forecasting setting should not be to produce correct imputations but rather to produce imputations which generate optimal predictions for the outcome of interest. In essence, it should not be considered problematic to use different imputation techniques for different aims even within the same study as imputation is not about re-creating the data which missing as accurately as possible. Rather, imputation is about re-creating the data which are missing such that the research objectives, whether these are predictive, inferential, or both, are affected as little as possible by the missing data.

One possible concern for the results of the paper is the choice of amputation strategy for generating missing values. The multivariate amputation used is based on a logit transformation of a linear model, which could give an undue performance advantage to the linear imputation technique (norm, reg, irmi, and amelia). However, due to the logit transformation of the weighted sum scores this bias should not be large, and the results themselves do not show any obvious biases in favor of these imputation strategies. A larger concern is perhaps that all amputations are made in the right tail of the resulting logistic distribution. This is especially a potential problem in relation to down-sampling whereby performance of a forecasting model may increase when data are excluded from the forecasting model. Evidence of this is clear in the results for the linear regression model when missingness is MAR and occurring in both the predictors and target (Figure 2). Generating missingness in a different part of the distribution could possibly yield different results for some of the simulations.

6 Conclusion

This paper has investigated the effects of missing data and imputation techniques on the predictive performance of random forest models and linear regression models when forecasting the number of fatalities from civil wars between 1989 and 2020. The results show that missing data poses a different problem in forecasting oriented studies compared to inferential studies, and that simpler single imputation techniques, except global mean imputation, seem to work well in order to maximize the predictive power. Just as in inferential studies, the problems with missing data increase with the proportion of missing values, the severity of the missingness mechanism from MCAR to MNAR, and when missingness occurs both among the predictors and the prediction target.

This paper also opens up a wider range of further inquiry into the effects of missing data for forecasting oriented studies. In this paper, the effects of different missing data handling techniques were assessed on simulated missing data under different conditions for missingness. Further studies should evaluate the effects of these imputation strategies on true missing data, and on data which are missing with mixtures of the three mechanisms. Additionally, this paper has focused on a continuous forecasting target with continuous predictors. As one potential issue with imputation is that you may have imputation which are outside the realm of possible values it would be valuable to repeat this study with classification as the main target. This would also allow for an evaluation of imputation techniques on order statistic performance metrics such as the AUROC and AUPR metrics which are commonly of interest in forecasting studies with binary outcomes. Further, the effects of different imputation techniques of variable importance scores in random forest models and other machine learning models need to be assessed as such metrics are often used to obtain inferences from forecasting models based on such algorithms.

References

Allison, Paul D (2001). *Missing data*. Sage publications.

Batista, Gustavo E.A.P.A. and Maria Carolina Monard (2003). "An analysis of four missing data treatment methods for supervised learning". In: *Applied Artificial Intelligence* 17.5-6, pp. 519–533.

Coppedge, Michael et al. (2021). *V-Dem Codebook v11.1*. Varieties of Democracy (V-Dem) Project.

Graham, John W. (2009). "Missing data analysis: Making it work in the real world". In: *Annual Review of Psychology* 60, pp. 549–576.

Hegre, Håvard and the ViEWS Team (2021). *Forecasting fatalities*. Typescript Uppsala University.

Hegre, Håvard et al. (2019). "ViEWS: A political violence early-warning system". In: *Journal of Peace Research* 56.2, pp. 155–174.

Hollenbach, Florian M. et al. (2021). "Multiple Imputation Using Gaussian Copulas". In: *Sociological Methods and Research* 50.3, pp. 1259–1283.

Honaker, James, Gary King, and Matthew Blackwell (2011). "Amelia II: A Program for Missing Data". In: *Journal of Statistical Software* 45.7, pp. 1–47.

Honaker, James et al. (2010). "What to Do about Missing Values in Time-Series Cross-Section Data". In: *American Journal of Political Science* 54.2, pp. 561–581.

Jadhav, Anil, Dhanya Pramod, and Krishnan Ramanathan (2019). "Comparison of Performance of Data Imputation Methods for Numeric Dataset". In: *Applied Artificial Intelligence* 33.10, pp. 913–933.

Japkowicz, Nathalie and Shaju Stephen (2002). "The class imbalance problem: A systematic study". In: *Intelligent Data Analysis* 6.5, pp. 429–449.

Kapoor, Sayash and Arvind Narayanan (2021). "(Ir)Reproducible Machine Learning: A Case Study". In:

Khan, Shahidul Islam and Abu Sayed Md Latiful Hoque (2020). "SICE: an improved missing data imputation technique". In: *Journal of Big Data* 7.1, pp. 1–21.

King, Gary et al. (2001). "Analyzing incomplete political science data: An alternative algorithm for multiple imputation". In: *American Political Science Review* 95.1, pp. 49–69.

Kowarik, Alexander and Matthias Templ (2016). "Imputation with the R package VIM". In: *Journal of Statistical Software* 74, pp. 1–16.

Kuhn, Max (2019). *3 Pre-Processing | The caret Package*. https://topepo.github.io/caret/pre-processing.html.

Lall, Ranjit (2016). "How multiple imputation makes a difference". In: *Political Analysis* 24.4, pp. 414–433.

Montgomery, Jacob M. et al. (2015). "An Informed Forensics Approach to Detecting Vote Irregularities". In: *Political Analysis* 23.4, pp. 488–505.

Muchlinski, David et al. (2016). "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data". In: *Political Analysis* 24.1, pp. 87–103.

Mueller, Hannes and Christopher Rauh (2018). "Reading between the lines: Prediction of political violence using newspaper text". In: *American Political Science Review* 112.2, pp. 358–375.

Pedregosa, F. et al. (2021). *Scikit-learn documentation: 6.4. Imputation of missing values*. https://scikit-learn.org/stable/modules/impute.html.

Pepinsky, Thomas B. (2018). "A Note on Listwise Deletion versus Multiple Imputation". In: *Political Analysis* 26.4, pp. 480–488.

Poulos, Jason and Rafael Valle (2018). "Missing Data Imputation for Supervised Learning". In: *Applied Artificial Intelligence* 32.2, pp. 186–196.

Robitzsch, Alexander and Simon Grund (2021). *miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'*. R package version 3.11-6.

Schouten, Rianne Margaretha, Peter Lugtig, and Gerko Vink (2018). "Generating missing values for simulation purposes: a multivariate amputation procedure". In: *Journal of Statistical Computation and Simulation* 88.15, pp. 2909–2930.

Schutte, Sebastian et al. (2021). "Climatic conditions are weak predictors of asylum migration". In: *Nature Communications* 12.1, pp. 1–10.

Shah, Anoop D. et al. (2014). “Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study”. In: *American Journal of Epidemiology* 179.6, pp. 764–774.

Shmueli, Galit (2010). “To explain or to predict?” In: *Statistical Science* 25.3, pp. 289–310.

Stekhoven, Daniel J. and Peter Bühlmann (2012). “Missforest-Non-parametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1, pp. 112–118.

Tang, Fei and Hemant Ishwaran (2017). “Random forest missing data algorithms”. In: *Statistical Analysis and Data Mining* 10.6, pp. 363–377.

Telenczuk, Maria (2022). *Imputing missing values before building an estimator*. https://scikit-learn.org/stable/auto_examples/impute/plot_missing_values.html.

Twala, Bhekisipho (2009). “An empirical comparison of techniques for handling incomplete data using decision trees”. In: *Applied Artificial Intelligence* 23.5, pp. 373–405.

van Buuren, Stef (2018). *Flexible Imputation of Missing Data, Second Edition*. CRC press.

van Buuren, Stef and Karin Groothuis-Oudshoorn (2011). “mice: Multivariate imputation by chained equations in R”. In: *Journal of Statistical Software* 45.3, pp. 1–67.

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke (2010). “The perils of policy by p-value: Predicting civil conflicts”. In: *Journal of Peace Research* 47.4, pp. 363–375.

World Bank (2021). *World development indicators 2021*. https://databank.worldbank.org/source/world-development-indicators.

Appendix

Table 3. Summary statistics of all predictor features and the predictor target

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
ln_ged_sb_target	0.422	1.277	0	0	0	0	11
ste_theta0	0.035	0.057	0	0.004	0.011	0.034	0.567
ste_theta2	0.014	0.025	0	0.003	0.006	0.016	0.415
ste_theta4	0.051	0.057	0	0.016	0.033	0.065	0.629
ste_theta5	0.033	0.038	0	0.012	0.022	0.040	0.551
ste_theta11	0.054	0.054	0	0.018	0.038	0.071	0.642
ste_theta13	0.048	0.091	0	0.007	0.017	0.044	0.939
ste_theta14	0.075	0.072	0	0.026	0.054	0.101	0.664
ste_theta0_stock	0.033	0.053	0	0.006	0.011	0.030	0.340
ste_theta2_stock	0.013	0.021	0	0.004	0.007	0.013	0.316
ste_theta4_stock	0.048	0.040	0	0.024	0.038	0.060	0.407
ste_theta5_stock	0.030	0.022	0	0.016	0.025	0.037	0.371
ste_theta11_stock	0.055	0.043	0	0.027	0.043	0.069	0.402
ste_theta13_stock	0.049	0.072	0	0.011	0.022	0.056	0.839
ste_theta14_stock	0.073	0.048	0	0.040	0.063	0.093	0.474
wdi_sl_tlf_totl_fe_zs	39.7	10.3	0	36.5	42.2	46.8	56.0
wdi_ms_mil_xpnd_gd_zs	2.337	3.016	0	1.082	1.695	2.685	117.4
wdi_dt_oda_odat_pc_zs	99	298	134	12	39	87	12,077
vdem_v2x_horacc	0.334	0.930	2	0.3	0.3	1.1	2
vdem_v2xnp_client	0.427	0.286	0	0.1	0.5	0.7	1
vdem_v2x_veracc	0.538	0.766	2	0	0.6	1.2	2
vdem_v2xcl_dmove	0.637	0.313	0	0.5	0.8	0.9	1
vdem_v2xpe_exlpol	0.347	0.295	0	0.04	0.3	0.6	1
vdem_v2x_divparctrl	0.009	0.947	2	0.7	0	0.8	2