An Evaluation of Machine Learning Imputation for Survey Data*

Steven Morgan, Xiaoran Sun, Claire Kelling, Shipi Kankane, Lulu Peng, So Young Park

November 18, 2018

Abstract

Missing data are a perennial problem for any social scientist analyzing survey data. If not addressed appropriately, missing data can bias inferences made from statistical analysis through a loss of valuable information or worse, result in selection bias. With declining response rates of survey respondents, missing data issues may now be exacerbated. Prior researchers have spent considerable time conceptualizing how missing data can bias analysis and have developed different methods to mitigate against the issues of missing data during the preprocessing stage of data analysis. Chief among these approaches is multiple imputation, a model-based approach to substituting missing values with estimates of their "true" value. Recent work outside the social sciences has begun employing machine learning algorithms to impute missing values, leveraging these models' increased predictive capacity to produce more accurate estimates of missing values. This machine learning-based approach to imputation represents an enticing approach to dealing with missingness in survey data. In this paper, we first implement multiple machine learning algorithms on a set of simulated data, and evaluate the algorithm's ability to recover simulated values under different states of missingness. We then implement these algorithms on a well-known and often-used survey dataset, the 2016 American National Election Study, and evaluate how different approaches to imputing missing data impact the inferences drawn from models of vote choice and candidate evaluation.

*This manuscript is a draft and was prepared for SODA 502.

Introduction

There often exists a discrepancy between how social scientists analyze data with missing values and how statisticians recommend dealing with missing data. While this gap has narrowed over the past couple of decades due to the proliferation of more efficient algorithms for missing data, chiefly multiple imputation and machine learning approaches, only the validity and bias of the former has been sufficiently addressed in the literature. Appropriate method to deal with missing data is important in terms of bias and causal inference [8, 4].

[18] articulated the different possible classifications of missing data. As [9] explicates, data can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Data are said to be MCAR when the probability of being missing is unrelated to the unknown true value of the variable in question, or to any other variable of interest. If missing data exist completely at random, estimates in a regression framework will not be biased, but should be less precise (due to a decreased N). The risk of false positives in hypothesis testing should not increase. More perilous implications arise when data are missing at random or not at random. Data are said to be missing at random when the probability of a missing response is independent of the value of the variable, conditional on other variables. If missing data are MCAR, then we can say that the mechanism of missingness is ignorable. In all other cases, missing data mechanism is said to be non-ignorable. Missing not at random brings with it the greatest potential for erroneous inferences. Missing not at random indicates that missing data in the dependent or one of the independent variables is systematically related to the measurement process or another variable, measured or unmeasured. However, MNAR in a variable cannot be predicted from the values of another variable. The risk of false positives increases compared to non-missing data in this case.

A principal example of missing data is in survey data. Missing data in surveys can take the form of unit non-response or item non-response. Possibly, five different sources of missing data on surveys might exist: 1) refusal to answer 2) incomplete responses 3) does not apply 4) respondents don't know 5) missing by design. There exists a vast literature on Total Survey Error, a framework that classifies a number of ways survey responses can be biased [2]. In terms of missing data, non-response is the measurement error of interest. If responses to a survey are missing completely at random, researchers need not be concerned with biased inferences. However, the survey methodology literature makes it quite clear that this is unlikely to occur. Item and instrument non-response are typically systematically related to a variable of interest. For example, item non-response is typically higher for questions on illicit drug use compared to items tapping age or education. This is problematic if illicit drug use is the concept studied and non-response is related to another concept of interest. While this is a rather extreme example where multiple imputation approaches may not be able to overcome this bias (i.e. a different methodology like a list survey should be employed), it makes clear that missing data can sufficiently bias estimates from quantitative analysis and inferences drawn. Traditional methods of handling non-ignorable mechanisms of missing data include list-wise deletions (complete case analyses), estimating the model with full information maximum likelihood model and imputation.

Machine learning approaches are becoming more common in social science research. In the medical field, machine learning has recently been utilized to impute missing data. Machine learning has proved useful for

these purposes for two reasons. First, these algorithms are scalable. Second, these methods utilize a rain/test split sample approach, allowing the analyst to cross-validate the results of the algorithm on a dataset. This latter approach represents an area of improvement over the current state of the literature on missing data. Multiple imputation approaches represent the current state-of-the-art in the social sciences. However, a major drawback is difficult interpretation and evaluation of the results of missing data imputation. The train/test split allows for the researcher to evaluate the performance of a machine learning approach on a dataset by holding out a proportion of the data. Cross-validation allows for re-sampling of a held-out proportion, which should be critically important for smaller survey datasets. This project will implement a variety of machine learning approaches to imputing survey data, taking into account unique challenges of imputing binary, categorical, ordinal, interval, and ratio values [3]. The effectiveness of these implementations will be measured by leveraging a train/test split as well as using k-fold cross-validation to assess model performance, a critical step not found in most implementations of multiple implementation.

Previous Approaches to Missing Data

There has often existed a gap between how social scientists have treated missing data in the pre-processing stage of survey data analysis and the literature on missing data contributed by statisticians and social science methodologists. The prior section touches on listwise deletion as an early approach in the social sciences to model data with missing values. Listwise deletion refers to the removal of an observation in its entirety if any value in that observation is missing across the variables modeled in a statistical analysis. For example, King et al. [8] conducted content analysis of all articles in the *American Political Science Review*, the *American Journal of Political Science*, and the *British Journal of Political Science* over a five year period and found that 94% of cases utilized listwise deletion. Further, they find that about one-third of data was removed on average. Even if all data were MNAR, this is a severe loss of statistical power due to the removal of information (less observations used for statistical modeling).

This is even more worrisome when data are MAR, where the loss of statistical power is overshadowed by potential bias existing in analysis. Biased inference refers to a situation in which statistical analysis results in different signs of estimated coefficients (the effect is in the opposite direction of the "true" relationship) or a difference in magnitude of estimated coefficients (the effect is smaller or larger than the "true" size of the effect). Biased inferences also refers to whether these inaccuracies occur in either a descriptive or causal framework (Anderson, Basilevsky, and Hum 1983). To refer back to King et al.'s analysis of articles published in top political science journals, the authors estimate that "the point estimate in the average political science article is about one standard error farther away from the truth because of listwise deletion" [8] (p. 52). In fact, the authors go as far as stating that the bias induced by listwise deletion is potentially hazardous that omitted variable bias in a model may even be preferable. An example would be as follows: if Democrats are more likely to produce item non-response to a question tapping propensity to vote, listwise deletion would bias inferences derived from

models of turnout by over-weighting the responses by non-Democrats. If NA values can be predicted by values in other variables (such as political ideology, income, race, education, etc.), the data are MAR. Imputation can provide estimates of missing values, boost statistical power, and importantly mitigate bias incurred by the underlying propensity for Democrats to skip this important survey item (selection bias). Conversely, if item non-response occurred at random, imputation could boost statistical power by increasing the number of observations modeled (compared to listwise deletion), but would be unlikely to help prevent bias. This occurs because if item non-response was random, it could not be predicted by values of other variables. It is important to note that when we say a variable can predict another variable, we do not mean in a causal sense. Instead, a variable can predict another variable if their exists a relationship between the distribution of variables (a correlation).

A novel approach that has been increasingly utilized over the last two decades is multiple imputation. This approach entails imputing m values for each missing item and creating m completed datasets. Across these now-complete case data sets, the original observed values are constant across the m datasets, but the original missing values are imputed with different values across the m datasets to reflect the level of uncertainty of the prediction of these values. Multiple imputation provides leverage by estimating multiple datasets, because the multiple imputations allow for evaluation of how stable these estimates are. Put more concretely, King et al. describes the value of producing multiple datasets as follows: "That is, for missing cells the model predicts well, variation across the imputations is small; for other cases, the variation may be larger, or asymmetric, to reflect whatever knowledge and level of certainty is available about the missing information" [8] (p. 53). This is a statistically valid approach, as set out by [17]: the "variance of the multiple imputation point estimate is the average of the estimated variances from within each completed data set, plus the sample variance in the point estimates across the data sets." Multiple imputation protects against the imputation of extreme values/outliers, by producing multiple sets of estimates (datasets); the researcher can evaluate the robustness of the imputation procedure. The approach assumes data are MAR, but imputing data that are really MNAR will not produce biased imputed values. While there is no set number for an adequate amount of datasets to impute, generally the researcher is encouraged to impute 5-10 dataset and evaluate the variance of imputed values. Ultimately, the researcher selects the dataset with the most valid imputations, and uses this dataset in statistical analysis. Multiple imputation can be calculated with an EM algorithm or Imputation-Posterior algorithm (based on Markov Chain Monte Carlo methods). Standard multiple imputation packages, such as Amelia II in R, utilize an altered EM algorithm (EMis: EM with importance resampling) due to its considerable speed upgrade. Additionally, IP algorithms are more computationally complex, as this approach calculates the full distribution, while EM is deterministic and calculates the maximum of the likelihood function.

Lastly, researchers have also dealt with the loss of statistical power as a consequence of missing data with more simple approaches. For example, missing values for continuous variables can be imputed by assigning the mean value of the distribution of non-missing values of the observation to the missing values. Similarly, for binary and ordinal variables, the researcher may take a similar approach but instead assign the median or modal

value of non-missing values to missing values. While this approach does increase the number of observations used for modeling, these approaches do not address, and can in fact worsen, bias induced from missing data. These approaches are less favorable approaches to dealing with missing data because they do not utilize information from the data beyond a single column (variable) to predict missing values. Additionally, if the mean value is a poor estimate for the missing values, this can induce bias in a different fashion than listwise deletion. To go back to the Democrat item non-response example, if the median value is skewed towards non-Democrats' responses, median- or mode-wise imputation will not mitigate against the over-weighting of non-Democrat responses. In fact, if there exists a substantial proportion of Democrat item non-response for this question, this approach could amplify this bias by treating these responses as more non-Democratic than the "true" values. A good imputation approach takes advantage of as much data as possible, under the assumption that MAR data can be predicted from non-missing data in the dataset.

Machine Learning for Imputing Missing Data

Multiple imputation represented an important leap in survey analysis because it was able to predict missing values using information from the data in a fairly accurate manner. While multiple imputation will not produce "perfect estimates", this is not an indictment of the method as much as an indictment of the stochastic element inherent in human behavior. Multiple imputation guards against poor estimates by producing multiple sets of estimates (imputations) and allowing the researcher to select the dataset that seems most reasonable. A potential shortcoming of multiple imputation lies in its underlying assumption that the variables are jointly multivariate Gaussian [11]. This is a substantial approximation, but research has found that more complicated assumptions of the underlying distribution of variables does not provide for improved validity of imputed values. This extends to categorical values as well. However, recent work has bypassed this assumption and instead turned to machine learning approaches to imputation that do not assume a multivariate normal distribution. For example, [6] utilize a variety of machine learning algorithms, specifically multi-layer perceptron, self-organization maps, and k-nearest neighbors (KNN), to impute missing data in a dataset of 3679 women with operable invasive breast cancer across 32 different hospitals in which about 5.61% of values were missing. They compared the predictive capacity of models predicting relapse with imputed data from these approaches to traditional mean-imputed and multiple imputation data. The models predicting relapse with imputed data from the machine learning approaches produced more accurate predictions compared to "standard" statistical approaches to imputation. Working under the assumption that the data are MAR (and thus can be predicted with other information in the dataset), relaxing the assumption of the joint distribution with machine learning approaches optimized for predictive accuracy produced more accurate predictions in the final modeling stage. Further, using two climate datasets, Richman et al. (2007) compared the performance between listwise deletion, linear regression, and two machine learning algorithms–Support Vector Regression (SVR) and Artificial Neural Networks (ANN), for imputation of missing data, and found SVR imputation with best performance, including the lowest mean squared error (MSE) and mean mean absolute error (MAE).

The intuition for machine learning imputation is similar to multiple imputation. The approach relies on assuming MAR data, and using as much information as possible to predict each missing value. For example, if values are missing in only one variable, the values should be imputed with a model that predicts that variable with missing data with every other variable on the right side of the model. In a more complicated (and realistic) case, two variables have missingness. All observations with missingness in only one variable should be predicted by a model with all other variables on the right side of the equation, and then observations with missingess in two variables should be predicted with a separate model with all but the two variables with missingess on the right side of the model. Machine learning imputation is thus an iterative procedure, where at each step as much data as possible (in terms of number of columns) are used to predict missing values [6]. The higher predictive capacity of machine learning algorithms should produce better estimates of missing values, especially when as much data as possible are used, similar to the standard multiple imputation approach. In this paper, we evaluate this statement with simulated and real survey data. These approaches have largely been utilized in the fields of medicine and meteorology. However, survey data of social and political behavior may differ drastically from the data sources used in these fields, both in terms of the rates of missingness and patterns of missing data due to the sometimes sensitive nature of polling individual behavior for the purpose of research. This project examines both the ability of machine learning approaches to recover values from simulated data and compares how data imputed with different approaches impact conclusions draw from statistical analysis of that data.

Analytics Approach

Across the simulated and real world datasets, this project will test two approaches to multiple imputation. First, missing data will be dealt with by two standard multiple imputation approaches, utilizing the 'Amelia II' and 'mice' package in R. Secondly, we will apply standard machine learning algorithms to impute missing data. We will utilize multiple algorithms, including SVM, Random Forest, ANN, KNN, and SVR. This allows us to test if algorithms that allow for hyper-parameterization with a grid search perform better than non-parametric approaches. The following sections briefly describe the approach utilized for imputation.

Multiple Imputation with Bootstrapped EM

Amelia II implements a bootstrapped-based EM algorithm of multiple imputation for the statistical software R. This package is extremely flexible, as it allows the researcher to impute cross-sectional, time series, and time series cross-sectional data with missing values. Additionally, the user is allowed to incorporate priors for this imputation procedure, though we do not set any priors in our implementation of multiple imputation with Amelia II. Categorical values are replaced with a series of dummy variables, and the distribution of dummy variables and continuous variables are assumed to be joint multivariate normal. The software package allows the researcher to select the number of imputed datasets returned, with a general guideline that 5-10 datasets

should be imputed and inspected. However, more than 10 datasets should be calculated when the proportion of missing data is exceedingly high (though what qualifies as a high proportion is not agreed upon in the literature).

Multiple Imputation by Chain Equations (MICE)

MICE uses built-in algorithms for multiple imputation of missing data, with a variety of imputation models applied to different types of variables, including continuous, binary, ordered and unordered categorical, and clustered continuous variables (Schouten,Vink, and Lugtig, 2016). In this study, we focus on the imputation of continuous, binary, and ordinal (i.e., ordered categorical) data. MICE uses predictive mean matching to impute continuous data, a method based on linear regression models that randomly samples among the closest matches in the predicted models with the non-missing cases. For binary and ordinal variables, MICE uses logistic regression and proportional odds regression, respectively, to impute missing data. For each dataset, we use mice to create 5 imputation datasets with 100 maximum iterations. Final models and results and pooled from results for the 5 imputed datasets.

Multiple Imputation using Artificial Neural Networks (ANN)

There has been previous attempts to apply Artificial Neural Networks (ANN) to imputations [13, 14, 12, 10, 7]. [13] developed an imputation measure that utilized ANN and [14] applied it to the 1990 Norwegian population census data. The imputation was carried out with 1,845 training data, about 10% of the original data, and feed-forward neural networks. He found the estimates to be reliable and unbiased. [10] implemented ANN models to impute the missing values of the National Resource Inventory (NRI) survey and to found it to outperform hot-deck and model-based imputations in regards to the mean squared error of prediction. [10] argue that application of ANN allows the large-scale data imputation to perform with minimum human intervention and less background information. More recently, [7] developed a deep neural networks based imputation method to impute race and ethnicity missingness in medical records. Their method, Race and Ethnicity Imputation from Disease history with Deep Learning (RIDDLE), generated more accurate and precise estimates for an anonymized medical records of 1,650,000 individuals from the New York City (Columbia University) and Chicago metropolitan populations (University of Chicago) compared to other imputation measures such as random forest and SVM[1]. RIDDLE is available as a Python package on Github (`https://github.com/jisungk/riddle`).

[1]Some references for ANN based imputations:

- Python: `https://github.com/jisungk/RIDDLE`
- Matlab:
  - `https://github.com/vishwakraam/Data-Mining-using-Aritifical-Neural-networks-and-SOM-`
  - `https://github.com/wxxia/NeuralNetwork_based_imputation`

Support Vector Machines

There is quite a bit of valuable research already in existence about missing data imputation through support vector machines (SVMs). For example, [22] discuss data imputation of activity-based transportation models. Specifically, they impute the number of cars in a home and the presence or absence of a drivers license through missing data imputation with an SVM. In this way, their analysis is similar to ours as they analyze their model under the case of both a binary response as well as a count response. However, their use of data imputation is also quite different than ours in a significant way. For their data, they assume they are only missing data in their response variables, the two listed above. This case actually may be quite easy to solve using a linear regression or glm, where they just need to predict the response using known covariates, although they do not compare their results to these models. In our case, we also have missingness in the covariate data, so it may prove to be a bit more difficult.

The authors of [15] also present an interesting example of using missing data imputation through SVMs. In this case, they also have missing data amongst the covariates, as in the case we are studying. They define a *risk* that is associated with using missing values when creating coefficient estimates. They create a case for minimal risk under squared loss with mean imputation. There paper is quite technical, but takes an interesting step towards quantifying risk of imputation using SVMs.

Lastly, [16] presents a recent Masters Thesis on using SVMs for classification and imputation. The first part of his thesis is quite standard, where he uses SVMs for classification through a typical dataset. However, the second part of his paper is quite useful for our study. He uses SVMs to impute missing categorical data and he demonstrates his technique on the 1997 National Labor Survey. He compares SVM to the Expectation Maximization (EM) algorithm and shows how it outperforms EM in the case of categorical variables and performs about the same in the continuous case. He imputes data without large attention to if it is a response variable or covariate. He also includes his code for imputation at the end of his thesis, found here.

Support Vector Regression

A regression framework that is based on Support Vector Regression (SVR) techniques is called Support Vector Regression[19]. Applying SVR techniques to imputing data have been attempted by some scholars. For example, authors in [21], employ SVR techniques to impute missing values in DNA micro-array gene expression data. As suggested by [5] and implemented by [20] , we employ SVR using three steps:

Loop until all attributes are imputed:

- Step 1: For a subset of the data that is complete and has no missingness, build models that map the relationship between a given column and the rest of the other columns.
- Step 2: Set one of input attributes, some of whose values are missing, as the decision attribute (output attribute)and the decision attributes as the condition attributes by contraries.
- Step 3: Using SVR, predict the decision attribute values.

The imputed values of attributes are combined as the model output. This is the imputed dataset. This algorithm uses the full information from each of the attributes to impute missing values. Owing to the sheer number of permutations that are carried out, applying SVR techniques to datasets with a significantly higher number of features can become burdensome.

Simulated Data

We conducted analysis on simulated data under various conditions to evaluate and compare the effectiveness of different approaches to imputing missing data, with the benefit of being able to compare imputed values across approaches with the underlying "true" value, treated as missing by each approach. Each dataset, D ,is drawn from a multivariate normal distribution, with mean 0 and variance 1, with n = 500 observations per dataset.[2] Drawing data from a multivariate normal distribution represents a stringent test of machine learning approaches to imputation compared to standard multiple imputation, because the standard multiple imputation approach assumes the data generating process is multivariate normal. In total, we present four different conditions of missing data. We vary the characteristics of missingness across the simulated datasets. Specifically, we vary the total missingess (number of cells) as well as whether the data are characterized as MNAR or MAR. Each dataset, D, contains five variables: X_1, X_2, X_3, X_4, and Y. To determine missingness, we simulate a second matrix M, with the same dimensions as the simulated data matrices, D, (500 rows, 5 columns). Each value of M is drawn from a uniform distribution spanning 0 to 1. We mirror the approach of King et al. (2001) in varying patterns of missingness; these four conditions are presented below:

- MCAR-1: Y, X_1, X_2, and X_4 are MCAR; X_3 is completely observed. For all values in M_{ij} (except for M_{i3} since X_3 is completely observed) greater than 0.97, the corresponding entry in the simulated data matrix, D_{ij} is set to missing. About 86.4% of the observations are fully observed.
- MCAR-2: $Y, X1, X2$, and X_4 are MCAR; X_3 is completely observed (same as MCAR-1). For all values in M_{ij} greater than 0.84 (except for M_{i3} since X_3 is completely observed), the corresponding entry in the simulated data matrix, D_{ij} is set to missing. About 50.8% of the observations are fully observed.
- MAR-1: Y and X_4 are MCAR; X_1 and X_2 are MAR, with missingness a function of X_3, which is completely observed. For X_1 and X_2, missingness is determined with the same approach from MCAR-1. If $X_{i3} < 0.8$ and $u, 0.98$, then M_{i3} or M_{i2} is set to missing. About 69.4% of the observations are fully observed.
- MAR-2: Y and X_4 are MCAR; X_1 and X_2 are MAR, with missingness a function of X_3, which is completely observed (same as MAR-1). If $X_{i3} < 0.98$ and $u < 0.1$, then M_{i2} and (separately) M_{i3} are also set to missing. About 50.6% of the observations are fully observed.

This approach allows us evaluate the efficacy of various approaches to missing data imputation under conditions that commonly confront researchers analyzing survey data. Since we simulate data before assigning

[2]The covariance matrix was specified as follows: (1 -0.12 -0.1 0.5 0.1, -0.12 1 0.1 -0.6 0.1, -0.1 0.1 1 -0.5 0.1, 0.5 -0.6 -0.5 1 0.1, 0.1 0.1 0.1 0.1 1)

missing values specified by the above four conditions, we can evaluate the performance of each imputation approach by comparing the imputed value against the original simulated value. We use root mean square error as our evaluation criteria, since the values produced in D are continuous. Root mean square error (RMSE) measure of the differences between values estimated by a model and the actual values of the data; the value represents the average error of the prediction across the dataset. Lower values indicate more accurate prediction estimates. Imputation approaches with lower RMSE values perform better at uncovering the true simulated values and should perform better on real survey data, conditional on the assumptions of missingness about the survey data are correct (mainly that the data are not MNAR). While survey data often contain non-continuous data (i.e. binary, ordinal, and categorical), we assert that this is an important first step to evaluating imputation performance. Further work will expand simulation tests to include other data categories that occur commonly in survey data.

Imputation of Simulated Data

This section analyzes the predictive capacity of the imputation approaches examined in this paper on the four simulated datasets. The below table displays the ability of the Support Vector approach to recovering the "true" values altered to missing values in the two MCAR and two MAR simulated datasets. The second table breaks out RMSE by variable (there are no RMSE values for X_3 across the four datasets because X_3 is fully observed in each, so no values of X_3 are imputed). Initially, it appears that SVR does a poor job across the datasets of predicting missing values. The size of the RMSE values is around 1.0 for the four datasets, and the values for each variable (excluding X_3) also hover right around 1.0. The RMSE values are not normalized, so the value represents the average error term for a predicted missing value compared to the actual simulated value. Remember that the values were drawn from a multivariate normal distribution with a mean of 0 and a variance of 1. That the RMSE values are about equal to the variance value is worrisome. Interestingly, the RMSE values do not vary substantially with the amount of missing data. Because SVR imputation relies on as much information as possible at each iteration to predict missing values, datasets with less missingness should produce more accurate predicted missing values. However, this does not appear to be the case here, even though the difference in proportion of missing values across MCAR-1 and MCAR-2 as well as MAR-1 and MAR-2 is about 30%. SVR does a slightly better job in MCAR-2 of recovering the "true" missing values, even though there are substantially more missing values. However, the opposite is true of MAR-1 and MAR-2: MAR-1 imputed values are more accurate than MAR-2. Additionally, the RMSE broken out by variable vary little within datasets. SVR appears to be a consistent estimator, even though the RMSE values are higher than expected.

Next, we evaluate the effectiveness of standard multiple imputation in recovering the true values of missing data across the four datasets. Using Amelia II, we run an imputation of m = 1 datasets for each dataset. Although the software allows the user to inspect multiple datasets, since this is a simulation approach of four datasets we do not want to bias results by advantageously selecting an imputed dataset. However, since the data

Table 1: RMSE for Support Vector Regression imputation of simulated data

Dataset	RMSE
MCAR1	0.943
MCAR2	0.934
MAR1	1.033
MAR2	1.087

Table 2: RMSE for Support Vector Regression imputation of simulated data by variable

Dataset	Y	X1	X2	X3	X4
MCAR1	0.944	0.951	0.961	-	0.938
MCAR2	0.934	0.934	0.934	-	0.950
MAR1	1.033	1.035	1.038	-	1.032
MAR 2	1.088	1.088	1.089	-	1.086

meet the underlying assumption of multiple imputation of multivariate normal and most of the data are not missing in each simulation, multiple imputed datasets should not vary drastically. Multiple imputation performs slightly worse than SVR. Each dataset has an RMSE value higher than produced by SVR. The RMSE values entail a substantial average error, similar to those produced by SVR. It is difficult to tell if this difference in values across SVR and multiple imputation is significant though, since the uncertainty of estimates of multiple imputation comes in the form of multiple datasets, not confidence intervals around the values themselves. Nevertheless, we take this as initial evidence that SVR performs similarly in terms of predictive capacity compared to multiple imputation implemented in Amelia II.

Table 3: RMSE for Multiple Imputation of simulated data (Amelia II)

Dataset	RMSE
MCAR1	1.118
MCAR2	1.100
MAR1	1.104
MAR2	1.241

We next evaluate how MICE recovers the missing values. The below table displays the resulting RMSE broken out by variable for the MCAR1 dataset. The MICE procedure does a substantially better job than SVR at recovering the "true" missing values, evidenced by the substantially smaller RMSE values. However, it should be noted that the predictive capacity of MICE does not appear to be as consistent across variables. For example, X_2 has a RMSE value less than half that of Y and X_4, despite the fact that they are drawn from the same distribution. Simply, there exists greater variation across RMSE values by variable than for SVR. However, this concern should be outweighed by the finding that the RMSE values are substantially lower than the values produced from implementing SVR.

Table 4: RMSE for MICE imputation with simulated data

	Y	X1	X2	X3	X4
MCAR1	0.2659	0.1290	0.1856	0.0000	0.2520

Imputation of ANES Data

As stated earlier, missing data is problematic for social scientists because it has the potential to bias inferences from statistical analysis. We analyze and compare how different imputation procedures impact the conclusions researchers draw from a standard regression analysis approach. We utilize the 2016 American National Election Study survey to compare how different approaches to missing data impact the results of statistical analysis. This dataset was selected due to its widespread use in the social sciences (ANES datasets have been cited over 7,000 times)[3] and its sufficient sample size (the 2016 dataset contains over 3,000 respondent observations). The ANES holds a special place historically in the social sciences, as its first iteration represented the first pilot study of political behavior of the US electorate. The ANES is a nationally-representative survey of voters. Respondents are sample before and after each general presidential election. We analyze survey data from their pre-election study. This decision was made to mitigate against potential bias inherent in responses tapping recall of a past event.[4] Importantly, this dataset allow us to select and run a well-specified and agreed-upon model from the voting behavior literature. It is important to use a well-specified model, because we want to ensure that results from models of different imputation procedures are not artifacts of poor model specification (i.e. high colinearity and non-robust estimates across models from slightly different data). This is especially important for survey data, because analyses of survey data often produce substantial unexplained variation in the dependent variable. With these potential hazards in mind, we draw on the US voting behavior literature to select two models to test across imputation procedures.

First, we model individual-level vote choice. While political scientists and election forecasters have had mixed success in predicting macro-level voting outcomes, researchers have had comparatively more success predicting individual-level vote choice.[5] Specifically, we model presidential vote choice (whether an individual reports she will vote Democrat for president or not) as a function of partisan identification, political ideology, and a host of control variables (race, sex, religiosity, education, and income) in line with the standard model specification from the political science literature [1]. We select this model to evaluate estimated regression coefficients, standard errors, and model diagnostics (AIC) from a logistic regression model, since we operationalize the dependent variable as binary (also in line with the extant literature). We analyze the impact on inferences when employing an ordinary least squares regression approach, modeling a feeling thermometer score of the Republican presidential candidate as a function of partisan identification, political ideology, and the control variables utilized in the vote choice model. We treat the feeling thermometer scores as continuous, in line with the prior literature, even though the values are bounded between 0 and 100 and may not perfectly fit the categorization of continuous.

[3] A full list of citations as of July 10, 2018 is available at https://electionstudies.org/wp-content/uploads/2018/07/ANES-Bib.pdf

[4] Responses to the item asking who the respondent *will* vote for president (pre-election) aggregate closer to the national popular vote in the actual election compared to the item asking who *did* the respondent vote for (post-election). Recall bias on this item is not random. Respondents systematically skew their responses in favor of the winner of the election.

[5] Ecological fallacy issues, difficulty in predicting turnout, and differing institutional electoral rules have all contributed to this interesting discrepancy/dilemma.

We make this decision to utilize a well-specified model informed by the vast literature on voting behavior, and defer to their operationalization decisions. As such, we utilize an OLS regression model. Similar to the logistic regression model of vote choice, we also measure the effects of different imputation procedures on inferences derived from the OLS model. Specifically, we compare estimated regression coefficients, standard errors, and model fit (R^2).

ANES Results

We first present the baseline models, using listwise deletion below. The below table presents the results of a logistic regression model of vote choice (Democrat versus non-Democrat). In line with the extant political science literature, partisan identification and political ideology are the best predictors of vote choice. Additionally, religiosity, education, and race/ethnicity are also important predictors. Income does not have a statistically significant effect, though this is most likely due to colinearity introduced into the model by simultaneously modeling education and income (they are strongly positively related). In terms of model diagnostics, it should first be noted that the ANES dataset contains 3,830 observations. Using listwise deletion on 11 variable reduces the dataset by over 1,500 observations. This should result in a substantial loss of statistical power. The model fits the data well, with an AIC value of 1,116.7.

We also present results of an OLS regression model of feeling thermometer values towards Donald Trump below. Similar to the model of vote choice, partisan identification and ideology are strong predictors of feeling thermometer responses about Trump, as are education, religiosity, and race. Additionally, income and gender have statistically significant effects as well. We lose slightly less observations in this model, in which 2,394 observations are used for calculating OLS. This is the case because more people do not respond to the vote choice question than the Trump feeling thermometer item, conditional on answering the items tapping the other measures across the two models. The model fits the data particularly well for survey data, with an R^2 value of 0.567.

Multiple Imputation with Bootstrapped EM

We now implement standard multiple imputation with Amelia II and run both models on the imputed datasets. Since multiple imputation treats binary variables as continuous, I transform the imputed values of each binary variable to a 0 if the value is less than 0.5 and 1 if not. The logistic regression model for vote choice is presented in Table 7. It is immediately clear that this model has more statistical power than the baseline model, due to modeling 3,830 observations. Politically informed is significant at the 0.05 level, and income is significant at the 0.10 level. This is also seen in that the standard error for each estimated coefficient is smaller than in the baseline model. None of the signs of coefficients flip. The only major change in coefficient size is for Democrat; the imputed dataset produces an estimated coefficient about 0.6 smaller than in the baseline model. The AIC of

Table 5: Logistic Regression Model of Vote Choice (Democrat vs. Non-Democrat)

	Dependent variable:
	Dem. Vote
Democrat	2.388^{***}
	(0.219)
Republican	-1.396^{***}
	(0.195)
Ideology	-1.069^{***}
	(0.077)
Bible Inerrant	0.330^{***}
	(0.120)
Pol. Informed	0.046
	(0.078)
Education	0.367^{***}
	(0.083)
Income	0.047
	(0.076)
Female	0.097
	(0.161)
Black	3.742^{***}
	(0.585)
Hispanic	1.991^{***}
	(0.282)
Constant	1.855^{***}
	(0.584)
Observations	2,228
Log Likelihood	−547.372
Akaike Inf. Crit.	1,116.745
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: OLS Model of Feeling Thermometer Scores of Trump (Values range from 0-100)

	Dependent variable:
	Trump Feeling Thermometer
Democrat	-17.638^{***}
	(1.307)
Republican	16.242^{***}
	(1.362)
Ideology	7.669^{***}
	(0.427)
Bible Inerrant	-3.917^{***}
	(0.773)
Pol. Informed	-1.136^{**}
	(0.473)
Education	-3.141^{***}
	(0.499)
Income	-1.340^{***}
	(0.454)
Female	-3.013^{***}
	(1.008)
Black	-10.940^{***}
	(1.996)
Hispanic	-10.519^{***}
	(1.737)
Constant	40.188^{***}
	(3.671)
Observations	2,394
R^2	0.567
Adjusted R^2	0.566
Residual Std. Error	23.813 (df = 2383)
F Statistic	312.585^{***} (df = 10; 2383)
Note:	*p<0.1; **p<0.05; ***p<0.01

this model is 2,633.7, a substantial improvement from the baseline model. More data produces a better model fit. In this instance, imputing data produced more precise estimates due to increased statistical power. Additionally, modeling different datasets produces different inferences on the role of how political informed individuals vote as well as the role of income.

Table 7: Logistic Regression Model of Vote Choice: Multiple Imputation Dataset

	Dependent variable:
	Democrat Vote
Democrat	1.665^{***}
	(0.126)
Republican	-1.213^{***}
	(0.132)
Ideology	-0.834^{***}
	(0.047)
Bible Inerrant	0.276^{***}
	(0.075)
Pol. Informed	0.147^{***}
	(0.047)
Education	0.252^{***}
	(0.051)
Income	0.084^{*}
	(0.047)
Female	0.076
	(0.102)
Black	2.796^{***}
	(0.264)
Hispanic	1.880^{***}
	(0.171)
Constant	1.126^{***}
	(0.354)
Observations	3,830
Log Likelihood	−1,305.839
Akaike Inf. Crit.	2,633.678
Note:	*p<0.1; **p<0.05; ***p<0.01

There do not exist substantial differences between the OLS model of Trump feeling thermometer values for the imputed dataset compared to the baseline model calculated with listwise deletion. The model of imputed data actually produces a worse-fitting model, as the R^2 from this model is 0.527 compared to 0.567 for the baseline model. This is a puzzling result, as in general more data produce better R^2 values.

Multiple Imputation by Chain Equations (MICE)

The logistic regression model of vote choice produces an estimated coefficient for Democrat identification smaller than the baseline model (2.338 versus 1.665), similar to the difference in beta's between the baseline and Amelia-imputed models. Politically informed now produces a statistically significant estimate (at the 0.01 level), com-

Table 8: OLS Model of Feeling Thermometer Scores of Trump: Multiple Imputation Dataset

	Dependent variable:
	Trump Feeling Thermometer
Democrat	-16.013^{***}
	(1.038)
Republican	16.378^{***}
	(1.122)
Ideology	7.588^{***}
	(0.346)
Bible Inerrant	-3.448^{***}
	(0.620)
Pol. Informed	-1.038^{***}
	(0.375)
Education	-3.581^{***}
	(0.403)
Income	-1.368^{***}
	(0.368)
Female	-2.365^{***}
	(0.823)
Black	-13.627^{***}
	(1.465)
Hispanic	-12.839^{***}
	(1.340)
Constant	40.872^{***}
	(2.964)
Observations	3,830
R^2	0.527
Adjusted R^2	0.526
Residual Std. Error	24.643 (df = 3819)
F Statistic	426.197^{***} (df = 10; 3819)
Note:	*p<0.1; **p<0.05; ***p<0.01

pared to failing to reject the null hypothesis in the baseline model. Income is significant at the 0.05 level, the strongest evidence for an effect of income from any of the models run. Interestingly, the estimated coefficient of female is significant at the 0.10 level. Both the baseline model and the Amelia imputed model produce no evidence of an effect of female on vote choice in the 2016 presidential election. The model fits the data well, with a pooled residual deviance value of 2,411.2.

Table 9: Logistic regression model predicting vote choice with MICE imputation

	estimate	std.error	statistic	df	p.value
(Intercept)	1.35	0.46	2.94	31.99	0.00
Democrat	1.92	0.16	12.00	48.47	0.00
Republican	-1.31	0.16	-8.15	56.17	0.00
Ideology	-0.92	0.09	-10.50	8.80	0.00
Bible Inerrant	0.26	0.10	2.60	29.09	0.01
Pol. Informed	0.16	0.07	2.46	20.07	0.01
Education	0.23	0.09	2.63	9.66	0.01
Income	0.14	0.06	2.18	21.88	0.03
Female	0.20	0.11	1.74	263.61	0.08
Black	3.65	0.42	8.73	34.52	0.00
Hispanic	1.82	0.20	9.11	66.18	0.00
Pooled residual deviance	2411.245				

Similar to the OLS model run on the ANES data imputed with Amelia II, this model with more data produces a worse-fitting model. The R^2 value for this model is 0.529. There do not exist substantial difference in terms of signs and significance of estimated coefficients between the baseline model and the MICE imputed model for Trump feeling thermometer values.

Table 10: Linear regression model predicting feeling for Trump with MICE imputation

	estimate	std.error	statistic	df	p.value
(Intercept)	40.74	3.35	12.16	64.05	0.00
Democrat	-16.09	1.08	-14.96	723.97	0.00
Republican	15.11	1.33	11.35	48.26	0.00
Ideology	7.78	0.40	19.40	56.31	0.00
Bible Inerrant	-3.93	0.70	-5.60	92.81	0.00
Pol. Informed	-0.93	0.39	-2.41	801.49	0.02
Education	-3.27	0.46	-7.17	81.36	0.00
Income	-1.38	0.46	-2.98	28.76	0.00
Female	-3.12	0.88	-3.54	209.65	0.00
Black	-13.92	1.60	-8.69	134.06	0.00
Hispanic	-13.37	1.48	-9.03	113.25	0.00
Pooled R-squared	0.5294				
Pooled adjusted R-squared	0.5282				

Discussion

In our simulation section, we find that one machine learning approach, support vector regression, slightly outperforms multiple imputation with bootstrapped EM, a standard approach to imputing data. This is especially surprising, because we simulated data under the assumptions made by multiple imputation, that the data are multivariate normal. We did not witness drastic differences in any model's ability to predict missing data based on the proportion of missing data or MCAR versus MAR situations. Machine learning with chained equations best recovered missing data under our the MCAR specification with 86% complete cases, and performed substantially better than multiple imputation with bootstrapped EM and SVR. However, MICE produced the greatest variation in RMSE across variables. Implementation on the other three conditions will further test the efficacy of MICE compared to SVR and standard MI.

We turned to the 2016 ANES (Pre-Election) to assess the impact on inferences drawn from statistical analysis of survey data. All implemented imputation methods allowed for greater statistical power than the baseline models run with listwise deletion, by increasing the number of observations drastically. However, model fit only improved across the logistic regression models of vote choice (for multiple imputation with bootstrapped EM and MICE), and not for the OLS models of Trump feeling thermometer values. The direction of the estimated coefficients was unchanged across all models; imputations procedures in these instances did not cause us to say an independent variable had an effect in the opposite direction compared to a model run on listwise deleted data. However, the size of the coefficient for Democrat identification changed substantially across the models when values were imputed. The baseline model produced the largest effect for Democrat affiliation. In terms of significance, different imputation approaches resulted in different interpretations of the effects of income and gender. Gender was only significant (at the 0.10 level) in the MICE imputed data, and income was significant in both the Amelia- and MICE-imputed models but not in the baseline model. Imputation procedures can bias whether or not a researcher finds an effect for important demographic variables in the study of political behavior.

Next Steps

Our next steps are to continue to implement the other machine learning models described in the first half of our paper in our simulation study and on ANES data. Due to the fact that machine learning imputation becomes a permutation problem when observations have missingess in multiple columns within a row, we have had to focus on coding only a couple of algorithms. We will continue to push forward with the others. Additionally, we will delved deeper into the reduction in model fit in the OLS regression models despite more data, especially if other imputation procedures produce similarly small R^2 values.

References

[1] Angus Campbell, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. *The American Voter*. University of Chicago Press, 1960.

[2] Don A Dillman, Jolene D Smyth, and Leah Melani. *Internet, mail, and mixed-mode surveys: the tailored design method*. Wiley & Sons Toronto, 2011.

[3] Andrew Gelman, Gary King, and Chuanhai Liu. Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443):846–857, 1998.

[4] John W Graham, Allison E Olchowski, and Tamika D Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, 8(3):206–213, 2007.

[5] Feng Honghai, Chen Guoshun, Yin Cheng, Yang Bingru, and Chen Yumei. A svm regression based approach to filling in missing values. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 581–587. Springer, 2005.

[6] Pedro J. Garcia-Laencina Emilio Alba Nuria Ribelles Jose M. Jerez, Ignacio Molina. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. 2012.

[7] Ji-Sung Kim, Xin Gao, and Andrey Rzhetsky. Riddle: Race and ethnicity imputation from disease history with deep learning. *PLoS computational biology*, 14(4):e1006106, 2018.

[8] Gary King, James Honaker, Anne Joseph, and Kenneth Scheve. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American political science review*, 95(1):49–69, 2001.

[9] Ranald R MacDonald. Missing data–quantitative applications in the social sciences. *British Journal of Mathematical & Statistical Psychology*, 55:193, 2002.

[10] Tapabrata Maiti, Curtis P Miller, and Pushpal K Mukhopadhyay. Neural network imputation: An experience with the national resources inventory survey. *Journal of agricultural, biological, and environmental statistics*, 13(3):255–269, 2008.

[11] James Honaker Matthew Blackwell and Gary King. A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research*, 2017.

[12] Giorgio E Montanari and M Giovanna Ranalli. Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472):1429–1442, 2005.

[13] Svein Nordbotten. Editing statistical records by neural networks. *Journal of Official Statistics*, 11(4): 391–411, 1995.

[14] Svein Nordbotten. Neural network imputation applied to the norwegian 1990 population census data. *Journal of Official Statistics*, 12(4):385–401, 1996.

[15] Kristiaan Pelckmans, Jos De Brabanter, Johan AK Suykens, and Bart De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, 2005.

[16] Spencer David Rogers. Support vector machines for classification and imputation. 2012.

[17] Donald Rubin. Multiple imputation for nonresponse in surveys. 1987.

[18] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[19] Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765, 1997.

[20] Qiang Shang, Zhaosheng Yang, Song Gao, and Derong Tan. An imputation method for missing traffic data based on fcm optimized by pso-svr. *Journal of Advanced Transportation*, 2018, 2018.

[21] Xian Wang, Ao Li, Zhaohui Jiang, and Huanqing Feng. Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC bioinformatics*, 7(1):32, 2006.

[22] Banghua Yang, Davy Janssens, Da Ruan, Tom Bellemans, and Geert Wets. A data imputation method with support vector machines for activity-based transportation models. In *Computational Intelligence for Traffic and Mobility*, pages 159–171. Springer, 2013.